# Deep learning pipeline reveals key moments in human embryonic development predictive of live birth in IVF

**Camilla Mapstone[1,2], Helen Hunter[3], Daniel Brison[3,4], Julia Handl[2], Berenika Plusa[1]**

[1]Faculty of Biology, Medicine and Health (FBMH), Division of Developmental Biology & Medicine, Michael Smith Building, University of Manchester, Manchester, United Kingdom

[2]Alliance Manchester Business School, University of Manchester, Manchester M15 6PB, UK

[3]Department of Reproductive Medicine, Old Saint Mary's Hospital, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK

[4]Division of Developmental Biology and Medicine, Maternal and Fetal Health Research Centre, School of Medical Sciences, Faculty of Biology Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK

## Abstract

Demand for IVF treatment is growing, however success rates remain low partly due to the difficulty in selecting the best embryo to be transferred. Current manual assessments are subjective and can lead to significant inter-operator variability. Deep learning techniques could lead to improved embryo assessment and live birth prediction, however previous attempts neglect early developmental stages and often require vast amounts of data. Here, we demonstrate that even with limited data it is possible to train convolutional neural networks to classify developmental stage at high accuracies and predict live birth from various time-points throughout development. We identify key windows that are optimal for assessing embryo viability and demonstrate the importance of incorporating information from earlier stages. Our outcome predictor models are competitive with, and potentially outperform, human expert selection. The pipeline produced here could lead to an improved, standardised approach to embryo selection compatible with multiple transfer strategies.

## Main Text

Machine learning (ML) has already demonstrated great promise in many areas of medical imaging[1-3] and has the potential to revolutionise the field of medical diagnostics. There has been recent interest in applying ML approaches to tackle infertility, a growing health crisis impacting both individuals and society that has led to a rising demand for in vitro fertilisation (IVF) treatments. Due to low success rates multiple IVF attempts are often required, leading to additional cost and distress for the patients. One of the main challenges in IVF is the difficulty in selecting the best embryo to be transferred, as at the moment there is no general consensus on exactly what healthy pre-implantation embryo development looks like.[4-6] The selection process routine in most clinics involves visual assessment of embryos in real time or via time-lapse videos. Embryos are assessed based on morphological features such as the number and size of cells at cleavage stage and in the trophectoderm (TE) and inner cell mass (ICM) in the blastocyst (Fig. 1A), the expansion of the cavity, developmental timings, cellular fragmentation, and multi-nucleation. This manual assessment is subjective, with up to 83% variation between embryologists[7], and uses only a fraction of the information potentially available. A machine learning (ML) approach that automatically assesses embryos using more extensive information from across the time-lapse videos could potentially provide a consistent and reproducible method for embryo selection[8]. However, existing ML approaches for embryo selection[8] still fall short in successful prediction of live birth, the lack of attention to pre-blastocyst stages, and the need for reproducible models trained on single clinic data.

Most previous attempts to train ML models to assess embryos from raw time-lapse data aim to classify embryos based on intermediate outcomes such as manual grading by embryologists or embryo aneuploidy, or early pregnancy outcomes after embryo transfer such as foetal heartbeat[7,9-21]. Only a few have been designed with the outcome of live birth[22-24]. Predicting live birth directly is a difficult task as there are a large variety of factors that contribute to treatment success, some related much more strongly to maternal factors than the embryo itself[25]. However, it is important to predict live birth as it is the overall aim of IVF and intermediate outcomes do not guarantee success.

Currently most ML studies solely focus on the blastocyst stage, with very few using earlier embryonic stages[11,14,17,24] and none making a live birth prediction on embryos before the blastocyst stage, mirroring the most widely used process of embryo selection by the trained embryologists. Although embryologists may pay attention to earlier stages, there is currently no standardised method of combining morphological features and/or assigned grades from across development into a single overall viability score. A few ML approaches have incorporated frames from early embryonic stages, but the frame selection based on video timestamp may be missing crucial moments or adding unnecessary data, potentially leading to overtraining. Identifying moments in development most predictive of live birth could lead towards an improved viability assessment, based on the most important information across development.

A significant barrier to the widespread adoption of ML methods in ART is the depth and breadth of data required to develop an algorithm sufficiently predictive to be of use clinically. Many previous attempts have used large aggregated datasets derived from a number of clinics over a number of years in order to generate sufficient data to train ML models[11,19,20]. This in turn is a major limitation to widespread adoption of ML in the sector, as most clinics have distinct treatment policies and patient populations[26] . There is a need for a versatile and robust ML algorithm for embryo selection that can be successfully tailored to the needs of each individual clinic.

In this work, we assessed the suitability of a convolutional neural network (CNN) for classifying images of early embryos. Firstly, through development of a stage classification algorithm we demonstrated the high efficiency of our chosen model at correctly classifying stages of pre-implantation embryos. We then showed that this ML algorithm can assess blastocyst viability at a similar level to highly trained embryologists and identified a limitation to blastocyst-based embryo selection. Subsequently, we successfully applied this model to predict live birth at earlier stages, for which no standardised assessment methods exist. Capitalising on that, we used ML to identify previously unknown pivotal time-points for live birth prediction in the pre-blastocyst embryo, which offers the means for better overall viability assessment and/or earlier transfer. Finally, we showed that combining information from our identified time-points of embryo development with blastocyst stage predictions allows for a quantitative ranking of high-quality blastocysts which could not have been achieved by any of the existing assessment methods. Critical to widespread clinical application, our findings also demonstrate that it is possible to train a ML model to a high standard using very limited data from a single clinic, paving the way for clinic specific, ML-assisted assessment of embryo quality.

## Results

All model training and testing was performed on a dataset of time-lapse videos of embryos with known transfer outcomes from the IVF clinic at the Department of Reproductive Medicine, St Marys'

hospital, Manchester, UK. This is an NHS-funded clinic with stable patient population demographics and standardised treatment policies, and therefore highly suitable for single centre treatment outcome studies[27]. Embryos were cultured in either Embryoscope ™ or Embryoscope+ ™ time-lapse incubator system (Vitrolife, Sweden).

**Developmental stage classification**

Currently, ML algorithms predicting live birth from routinely collected time-lapse videos have had limited success, however it is unclear whether this is because the models are not well suited to the data, or if it is due to the inherent difficulty of live birth prediction after IVF procedures. Therefore, we first investigated how well our chosen model worked on our dataset for the relatively straightforward task of developmental stage classification, before attempting to train the model for the more challenging problem of live birth prediction.

To develop stage classification models, all time-lapse videos had one frame extracted for each developmental stage, resulting in equal class sizes. To reduce the amount of training data needed, we adopted a transfer learning approach, using the MobileNetV2 model with layers pre-trained on ImageNet, the overall workflow is illustrated in Fig. 1B.

First, we trained a binary model to classify embryos as zygote (one cell stage embryo that is a product of fertilisation) or 2-cell stage (an embryo after the first embryonic division). Subsequently, we trained a second binary model to recognise subtler differences on the subcellular level. To this end, we used images of embryos taken before or after nuclear envelope breakdown (NEBD), a process that marks the transition from interphase to the first embryonic division. The two investigated models were trained with varying amounts of training data by randomly excluding a portion of our data set. The training set size ranged from 10 to 1200 (all data used) with and without data augmentation. The validation and test sets were kept constant at 100 images each. The average test set scores (Fig. 1C-D) showed that when all the data was used we reached an accuracy of 97.1% for the zygote vs 2-cell model and 94.6% for the before NEBD vs. after NEBD model.

For both models the accuracy increased with the amount of training data, as expected, and appeared to be reaching a plateau in performance once the number of images in the training set rose to around 200-400. This plateau may suggest that the model performed close to its optimal level and we are not likely to gain much better performance by adding more training data. Visual inspection of the images (Fig. s1A) also suggested that there is an upper limit to the accuracy that could be obtained as vacuoles and cells dividing in a plane that does not allow subsequent blastomeres to be immediately identified can cause an embryo to appear to be at a different stage when viewed without the context of the time-lapse video.

Our results also showed that when using pre-trained layers, relatively high performances can be achieved even with very little data, over 85% accuracy was obtained for both models with a training set size of 10 (5 images from each stage). Fig. 1C-D also show that augmentation (see methods for details) seemed to have a small but positive effect, especially with limited training data, so we chose to continue using augmentation going forward.

We then trained multi-class models to classify images into five output classes corresponding to the consecutive stages of human pre-implantation development (Fig. 1E). We saw that an accuracy of 87.7% was achieved, which is much higher than the by-chance score of 20%. The results in Fig. 1E

show a similar trend to the one observed in the binary models; increases in accuracy became small once the training set reached about 200-400 images, and a high accuracy (69.4%) was achieved with a training set of just 10. These results further demonstrate that MobileNetV2 with pre-trained layers can achieve a high performance, even with a small amount of the training data, when analysing routinely collected images of preimplantation human embryos.

**Live birth prediction at blastocyst stage**

Following the successful development of embryonic stage classifiers, we next trained the model to predict live birth outcome based on the blastocyst stage, as this stage is most commonly used for embryo assessment. We extracted the final frame from each time-lapse video and trained the model five times (with 50 repeat training runs) using five mutually exclusive test sets that spanned the whole dataset when combined. For each embryo, the 50 resulting model confidence scores from the 50 training iterations were averaged to produce a final prediction for that embryo. We then calculated the ROC AUC; the area under a curve (called the ROC curve) that is created by plotting true positive rate vs. false positive rate at various thresholds. A ROC AUC of 0.5 is no better than chance and 1 is a perfect model, we achieved a ROC AUC of 0.680 for live birth prediction at blastocyst stage.

Next, we wished to compare our results against embryologist selection performance. To do this we used a subset of embryos for which Gardner grades[28] were available (141 in total) to produce ROC curves using the average test set scores and embryologist grades. In order to produce the ROC curve from embryologist grades we first converted the Gardner score letter grades assigned to the ICM and TE into numbers, as proposed by Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology[29]. Then an overall score was obtained by calculating an average of TE, ICM and expansion scores.

The ROC curves for both blastocyst assessment methods are shown in Fig. 2. The ROC AUC was similar for embryologist and model assessment, 0.720 and 0.726 respectively. This suggests our model had a very similar performance to highly trained embryologist grading, as St Marys' clinic host the UK NEQAS (a charitable consortium of external quality assessment laboratories) in reproductive biology. It is, however, important to note that the blastocyst model was at a disadvantage due to the use of limited information; only a single focal plane from just the final frame was used to train the ML model. Whereas the assessment of the embryologist was done using multiple focal points from the final frame, with possible adjustment of the score according to the time-kinetic data.

To investigate the effect of using just one focal frame for each embryo, we examined the confidence scores given to embryos where the ICM was not visible but the TE appeared to be good quality (Fig. S2A). We found that these embryos tended to receive very high confidence scores, suggesting that the ML model may not have used the ICM to make predictions. We then tested this further using the LIME software[30], which illustrates which parts of an image most strongly influenced a model classifying that image. The explanation images produced (Fig. S2A) appeared to focus on the TE, possibly using the size of the cavity to determine blastocyst quality. The ICM was often not included in the areas highlighted as useful to the model at all, providing further evidence that the model may not have been using ICM data.

In clinical practice, a numerical scoring system of embryo viability is more useful than a binary 'successful' or 'unsuccessful' prediction, as it allows the embryos in a cohort to be ranked in order of

likelihood of live birth. Although our model was trained as a binary classifier, the confidence score is continuous so potentially could be used to rank embryos. We investigated this by sorting all the embryos in our dataset into buckets by average test set model score, bucket 1 being the lowest score and bucket 10 being the highest score. The success rate (number of successful embryos in bucket/total number of embryos in bucket) for each bucket is shown in Fig. 2B. There was a general increase in success rate from bucket 1 to 10, suggesting that the confidence score was correlated to chance of successful transfer and therefore could be useful as a method to rank embryos in a cohort. However, increase in success rate with bucket number appeared to plateau from bucket 7-10, suggesting that there is a subset of high quality blastocysts for which any further increase in model score has little correlation to increased chance of live birth. Analysing embryologist grade versus live birth for a dataset of fresh blastocyst transfers from St Marys we saw a similar plateau (Fig. S2B), suggesting that this is an issue for both ML and embryologist based approaches.

Our data demonstrated that it is possible to train MobileNetv2 to predict live birth from blastocyst morphology to a performance similar to embryologist assessment, with potential for further improvements to performance. Importantly, using both the ML and embryologist scores we have identified a limitation to embryo assessment based on blastocyst images; above a certain quality threshold it appears to be difficult to discriminate further the most viable embryo from a group.

**Live birth prediction across development**

Next, we wanted to investigate the possibility of training models on pre-blastocyst (zygote, cleavage and morula) stages as this could allow for information from across development to be included in embryo assessment and also could potentially assist earlier selection and transfer in the future. Therefore, we extracted frames at four other precisely defined stages; one hour before NEBD completion, first appearance of 2 cells, first appearance of 4 cells, and initiation of 8-16 cell division round, we referred to these stages as PN, 2-cell, 4-cell, and 8-16 cell respectively. Live birth prediction models were trained using each stage.

As predicting live birth from embryo morphology alone is a difficult task, we decided to use our multi-class stage classification model as an extra transfer learning step (see methods section for full details). We experimented with various hyper-parameters, the results for all stages are shown in Fig. S3A. Our results suggest that this extra transfer learning step generally improved model performance, particularly for the PN and 4-cell models.

The ROC AUC values reported in Fig. S3A were the average of 50 training iterations each with a different randomly selected test set, as we had observed that the training uncertainty was minor compared to changes in test set. This also allowed for a quicker investigation than the five mutually exclusive test sets with 50 training attempts used in the previous section. We chose to use varying test sets to increase statistical power as we found that separate test sets had a higher than desirable variability in performance, reflecting the heterogeneity of the data in this ML task (Fig. S3b). However, the transfer model was trained on the same dataset, so to check this was not resulting in an unfair bias it was necessary to also develop all models from scratch with a true 'hold out test set' – a test set separated out at the very start, prior to training the transfer model. As shown in table 1, our results continue to hold up in this control experiment, so going forward we use transfer learning (with our optimised hyper-parameters) for the PN and 4-cell models and the original MobilenetV2 model for all other stages.

The average test set ROC AUC over 50 training attempts with random train/validation/test split is shown in Fig. 3B for each stage. The blastocyst stage has the most obvious morphological differences between classes, with unsuccessful embryos less likely to form expanded blastocysts, and as expected this model had the best performance. However, the pre-blastocyst models still gave above chance predictions. From our model performance it was clear that different developmental stages carry different predictive power. To identify specific moments in development for which life birth predictions are the most successful we tested multiple frames at regular intervals using each of the previous developmental stages as reference points (Fig. 3C-G),

We found that the model performance varied at different time-points and appeared to peak at certain moments in development as shown in graphs 3C-G. For the PN stage (Fig. 3C), a performance peak was observed just before PNBD. Performance across the 2-cell stage (Fig. 3D) was variable with no peak seen. A performance peak was observed 14 hours after 4-cell (Fig. 3E), this generally is within the 4-8 cell transition however no potential explanation for this peak was found from visually examining the time-lapse videos. Another performance peak was observed 21 hours after initiation of 8-16 cell cycle (Fig. 3F), when embryos tend to be in the morula stage. As an initial examination of the time-lapse videos suggested that this was generally just before cavitation, we then quantified this by counting the number of cavitating embryos at this time-point and at time-points just before and after. The results, shown in table 2, confirmed that this performance peak corresponds to the moment just before cavitation. Lastly, the blastocyst stage showed a gradual improvement in model performance towards the end of the time-lapse video (Fig. 3G), appearing to reach a plateau 6 hours before the last frame, by which point the successful embryos have usually formed an expanded blastocyst. We found that the peaks at PN, 4-cell +14hrs and 8-16 cell +21hrs were all statistically significant when compared to a time-point 6 hours earlier (p values of 0.0001, 0.0005, and 0.0011 respectively).

Finally, to test the potential benefit of averaging over repeat training attempts and also obtain a single prediction for each embryo in our dataset (to be used in further investigations reported below), we re-trained each model using five mutually exclusive test sets and 50 repeat training runs for each test set and calculated an 'average score' for each embryo from averaging over these 50 training iterations (as in the previous section). The PN, 2-cell and blastocyst stage models were trained using the original time-point and the 4-cell and 8-16 cell stage models were trained using the peak performance time-point (4-cell +14hr and 8-16 cell +21hr). The ROC AUC scores calculated from these averaged scores (table 3) showed that slightly better predictions can be made when the confidence score assigned to an embryo is averaged over many separately trained models, which is consistent with the existing literature on ensemble learning[31]. Additionally, we then also used the embryo average scores to calculate F1 and precision-recall AUC for each model (table S1) for further validation.

**Combined model outputs allow refined ranking of high grade blastocysts**

We next wanted to investigate whether our pre-blastocyst models could assist with the previously unreported issue we identified earlier – the difficulty in further differentiating between embryos above a certain blastocyst quality threshold. To test this, we calculated the ROC AUC of models at each stage on this subset of 'high quality' embryos. The results (shown in Fig. 4A) show that the blastocyst model performed almost no better than chance and worse than all the earlier stage models, and the PN model performed the best. This suggests that if multiple embryos in a cohort fall

into this high quality blastocyst sub-group (blastocyst model score>0.83), then pre-blastocyst stage models should be used to choose which one of these high quality blastocysts to transfer, rather than simply choosing the one with the highest blastocyst score.

To further investigate the benefit of using earlier stage models in conjunction with the blastocyst model we plotted the pre-blastocyst score vs. blastocyst score for each embryo, as shown in Fig. 4B. These graphs provide a visualisation of the spread of scores assigned by each model and the correlation between earlier stage scores and blastocyst score. We noticed that the model confidence scores had a tendency to be close to the extreme values; 0 or 1. To further demonstrate the added value of the pre-blastocyst stage models we calculated the success rate of our 'high quality' blastocysts when the pre-blastocyst model score was >0.9 vs <0.1. The results, displayed on Fig. 4B, showed that within the 'high quality blastocyst' group, embryos with a very high (>0.9) early (PN and 2-cell stage) model score were substantially more likely (54 v 34% and 47 v 21%, respectively) to be successful than the embryos with a very low (<0.1) early model score. This adds further evidence that pre-blastocyst model predictions should be taken into account when selecting from high quality blastocysts.

**Discussion**

IVF success rates are currently restricted by the difficulty in selecting the best embryo for transfer, and developing ML algorithms to assist with this task is an active area of research[8]. This study is the first to use early stages of pre-implantation human development prior to the blastocyst to predict live birth outcome following transfer, and is also the first to demonstrate these pre-blastocyst predictions can be used to refine the selection for transfer of high quality blastocysts. Our pipeline is reproducible and can inform multiple transfer strategies, therefore could lead to a universally applied standard protocol for embryo selection.

Prior to predicting live birth, we wanted to investigate whether the model was suitable for assessing embryo images. Developmental stage classification is relatively straightforward, while live birth prediction is much more difficult as some embryos fail due to maternal factors rather than the embryo itself. This essentially adds 'noise' to the data; some 'unsuccessful' embryos will actually be of a high quality and likely developmentally competent. The high performances achieved in stage classification demonstrates the high capability of our chosen model, even for quite subtle subcellular differences such as classifying embryos as before NEBD vs. after NEBD. This provides evidence that difficulties encountered when predicting live birth are not just due to deficiencies in the modelling system or the model being unsuited to the dataset, as it performed very well when there was no overlap between classes.

There have been many ML studies assessing embryo quality in recent years, however only a few have attempted to directly predict live birth (using a dataset exclusively containing transferred embryos). Two studies have attempted to predict live birth using images extracted from timelapse at various time-points; Ueno et al.[23] and Sawada et al.[24]. Ueno et al. used the deep learning model IVY, which had been developed by Tran et al.[11] to predict fetal heartbeat using 10638 embryos. Using IDAscore, an algorithm based on IVY, Ueno et al. attempted to predict live birth and reported a ROC AUC varying from 0.66 for age 41-42 to 0.76 for age>42. No average score for all ages was given so it is not possible to give a direct comparison, however our highest performing model; - the blastocyst model, achieved within this range (0.69). Sawada et al. achieved a ROC AUC of 0.64, lower than our

blastocyst model, however it should be noted they had a very limited amount of data available (only 91 successful embryos). Miyagi et al.[22] predict live birth from just blastocyst images, so is the most directly comparable to our blastocyst model. However, they also only report ROC AUC by age, with the value ranging from 0.634 for under 35 to 0.866 for over 42. No overall result is given, but again our blastocyst model falls within this range.

We found that the model performed at least as well as embryologist grades at predicting live birth at the blastocyst stage and it could be possible that performance may improve further if more temporal and spatial information were to be included. Our ML blastocyst model did not appear to be using the ICM to make predictions, possibly due to the fact that the ICM would have been out of focus in some of the training data. It is possible that including frames where the ICM is seen more clearly or incorporating a model specialised to the ICM could lead to a higher performance for the blastocyst model, allowing it to significantly outperform embryologist assessment.

Currently, most embryo assessment approaches are focused on the blastocyst stage, especially when using ML algorithms. Here, we have shown it is possible to predict live birth from stages earlier than the blastocyst, with ROC AUC values of 0.578-0.603. These pre-blastocyst predictions could be useful for various transfer strategies; they could either be used to improve selection at blastocyst stage, or they could assist selection at an earlier stage, allowing for more flexible treatment regimes. We have provided evidence that predictions from early development may be used in conjunction with predictions from the blastocyst stage to give a better assessment of embryo quality than the blastocyst model alone. Further investigation into combining multiple model outputs and optimising decision thresholds would be an exciting avenue for future studies.

Embryo assessment, both manual and ML, typically has an emphasis on specific chosen timepoints, however these may not necessarily be the optimal moments of embryo development for assessing viability. Here we identified the key developmental time-points that produced the best live birth predictions. We noticed that the peaks in model performance found at PN stage and 8-16 cell +21hrs are both just before certain well described developmental events; NEBD and blastocyst cavitation respectively. There are two possible explanations for this; at these pre-event time-points there is less natural variation in the appearance of viable embryos (compared to time-points where the embryo is undergoing processes such as PN growth or compaction) so it is easier to distinguish important developmental abnormalities, or these may be biologically important moments where any deviation from normal development can prevent the embryo from developing properly. For example, the ability of embryos to correctly prepare and execute the first mitosis is one of the defining moments of development and any anomalies around this time, such as problems with the NEBD process, can result in the developmental failure. The peak during the transition from the 4 to 8 cell stage coincides with embryonic genome activation in human embryos[32,33], raising the possibility that there are some morphological manifestations that can indicate the successful activation of the genome.

Additionally, our findings could be used to support earlier transfers. Extending the culture until blastocyst stage is associated in some studies with a number of adverse outcomes from pregnancy and early childhood, including pre-term birth, altered birthweight, and mono-zygotic twinning[34-37]. It is also possible that blastocyst culture may have negative implications on healthy aging, as recently it has been linked to shortened telomeres[38]. These disadvantages are not unexpected, since blastocyst culture acts to expose embryos to selection stress in a non-physiological in vitro environment, at the

precise time in development when the embryonic genome and epigenome are being reset as part of the formation of a new individual[4,38,39]. Blastocyst culture is currently favoured as it leads to significantly higher live birth rate, per transfer and per embryo[40], however the cumulative live birth rate, including transfer of all fresh and frozen embryos from a single egg collection event, are not necessarily increased using this policy. Our pipeline provides the flexibility of making viability assessments at various points in development, supporting various clinical strategies.

One of the biggest common limitations to ML is large data requirements. However, the models trained here demonstrate the diagnostic power of ML in IVF procedures without needing large amounts of training data. Our stage classification models showed that increases in performance became small when training set size was increased beyond around 200-400. This could suggest that although our outcome models might marginally benefit from more data (our smaller training set class had less than 200 embryos), the data amount was not likely to be the main limiting factor in performance. Our live birth prediction models were trained on relatively small number of embryos compared to many other studies, suggesting that our algorithms could easily be re-trained on an individual clinics' patient population to become specifically tailored to that clinic.

Our stage classification models also showed that MobileNetV2 pre-trained on ImageNet can give high accuracies on our dataset even with very small amounts of training data for simple classification tasks. Other studies have also used ImageNet for classifying human embryos[7,9,10,12,13,17-19,21,24], however our work is the first to investigate the performance that can be achieved with very little embryo data. Our data is quite different from the typical images included in ImageNet, which is mainly composed of photographs of everyday objects and animals rather than medical images. Therefore, it is useful to see that the fixed high level features obtained from ImageNet can be used successfully in this context.

We acknowledge that classification problems involving datasets with a lot more variation in background/magnification etc. may need substantially more data, however our results here demonstrate that, under certain conditions, deep learning does not necessarily require huge datasets. This could be of interest for a wide range of biological and medical applications where data is severely limited.

**Conclusion**

In summary, we have trained an ML model to both classify the developmental stage of an embryo and predict live birth using single-clinic data. We have also identified specific windows of early development that are most predictive of transfer outcome. The findings presented here are the first to predict live birth before the blastocyst stage and we have provided evidence that these pre-blastocyst predictions could be combined with blastocyst stage predictions to give better overall embryo assessment than selecting based solely on blastocyst morphology.

## Material and methods

### 1. Patients and time-lapse videos

Time-lapse videos of developing embryos were supplied by the IVF clinic in the Department of Reproductive Medicine at Old Saint Mary's hospital, Manchester University NHS Foundation Trust, Manchester, UK. Embryos were cultured in either Embryoscope ™ or Embryoscope+ ™ time-lapse incubator system (Vitrolife, Sweden). Throughout the study period there were no changes to the

culture media or other culture conditions. Embryos were cultured in GTL overlaid with Ovoil (Vitrolife, Sweden) from post-injection (after ICSI, approximately 40 hours post-hCG trigger) to day 5 of development. Embryoscope ™ slides hold up to 12 embryos in individual wells holding 25µL of media, overlaid with 1.2ml oil. Embryoscope+ ™ slides hold up to 16 embryos, in 2 rows of 8 wells, each row of 8 wells is overlaid with 180uL of media. Embryos were not removed from the incubators during the observation periods, and media was not changed. Each time-lapse video exported had a framerate of 5-10 frames an hour. All images were irreversibly anonymised by clinic staff before being given to us.

The dataset comprised of fresh ICSI transfers from 2016-2019 that resulted in either live birth or no pregnancy. Both single embryo transfers (SET) and double embryo transfers (DET) resulting in either no pregnancy or male/female twins (to exclude the possibility of monozygotic twinning) were included. In total we used time-lapse videos for 443 successful embryos and 257 unsuccessful embryos.

**2. Preparation of input image data**

The frame number of specific stages in development were recorded by viewing each video in ImageJ. The frames at various time intervals before and after this moment were then automatically extracted by training a CNN to read the timestamp. Measuring time in hours rather than frames was neccessary as the time between frames was inconsistent between videos.

Where the image quality was too low or the specific moment could not be determined (due to out of focus cell divisions or excessive fragmentation) the embryo was not included for that stage. We ensured the Embryoscope/Embryoscope plus ratio was equal across the two groups by randomly removing some of the successful Embryoscope embryos (this group initially had more embryos recorded by the EmbryoScope).

All images before the blastocyst stage were cropped to 300x300 pixels as this was the smallest size that captured the whole embryo including the zona. As embryos at the blastocyst stage expand a variable amount, occasionally almost filling the image, this stage was left uncropped. All images were then resized to 224x224 as this is the input size required by the pre-trained model.

**3. Model**

The model we have used is the MobilenetV2 model[41] with weights pre-trained on the ImageNet database[42]. We used fixed convolutional features, only training the final layer of the model. All developmental stage and initial transfer outcome training attempts used a base learning rate (BLR) of 0.0001, a drop out of 0.5 and the cross entropy loss function. For the transfer outcome models we applied a class weight of 2 to the live birth class to account for this class having about half the amount of data as the no pregnancy class.

For the transfer outcome models we also experimented with using multi-class stage classification for extra transfer learning. The first step was to add an extra hidden layer to the MobileNetV2 model and train it to predict developmental stage with both the hidden layer and last layer trainable (the rest of the network was fixed as before). In the 2nd step of training we then took this model as a starting point for our outcome prediction models, for this step the only trainable layer was the last layer, so all the convolutional layers and the hidden layer were fixed. We repeated these two steps with various numbers of hidden units (100,320,640, or 960) in the hidden layer to find the optimal

model structure. For stages earlier than the blastocyst we used the classes as in Fig. 4E for step 1. For the blastocyst stage we trained a separate model with the before NEBD class replaced with a blastocyst class and all images uncropped so that the resolution was the same at all stages.

**4. Model training**

For the training attempts where we had a hold out test set we had 50 successful embryos and 96 unsuccessful embryos in the test set and validation set (the same proportion as in the dataset as a whole). For all other training attempts we had 25 successful embryos in the test set and validation set and 48 unsuccessful embryos. For all training attempts we ran 50 iterations, we randomly assigned embryos to the training, validation and test sets at the start of each iteration (where a hold out test set was used just the validation and training sets were randomly assigned). When assigning embryos to the test and validation sets we kept embryos from the same cycle together for the transfer outcome models as they had the same camera settings so could give an unfair bias if split between the training and test set. To do this we had lists of all double transfers and single transfers in each class (successful/unsuccessful) and randomly picked embryos from these lists, keeping the ratio of double to single transfers the same as in the original dataset. We then performed augmentation on the training set by rotating each image by 90,180 and 270 degrees and getting the mirror image.

Each training attempt was repeated twice, the first time we ran each model for 10,000 epochs and recorded at what epoch the validation set peaked in performance on average across all iterations. The second time we trained the model for this optimal number of epochs every iteration. We did this because our models were prone to overfitting so finding the optimal number of epochs was important, however the small size of our validation set made it unreliable for stopping the model at the place that would also be optimal for the test set. We found that generally we got slightly better performance on the 2$^{nd}$ training attempt when we were using the optimal number of epochs.

**5. Model performance evaluation**

For each model setup we ran 50 training iterations and for each iteration the area under the curve (AUC) of the receiver operator characteristic (ROC) was calculated. The ROC is a graphical plot of true positive rate vs. false positive rate at various thresholds. The AUC of this curve is therefore a measure of how well the model performs, 1 corresponding to a perfect classifier. We calculated the ROC AUC using Scikit-learn python library. The standard error in ROC AUC values all iterations was then used to calculate error bar values.

The blastocyst model was compared to the embryologist scoring system using grades assigned by the embryologists at St Marys. An overall 'embryologist score' was calculated from averaging TE, ICM and expansion scores, we acknowledge this scoring system is imperfect as in reality the embryologist selecting the embryo may be able to select the better embryo out of embryos given the same grade and may place a different level of importance on the individual expansion/TE/ICM scores. However, due to the lack of an evidence base for selecting amongst embryos of similar grades we decided a simple average would be best for calculating an overall score, allowing for a comparison with our blastocyst model.

For the blastocyst model, we used the LIME software[30] so investigate how the model was choosing to classify blastocyst images. LIME (short for Local Interpretable Model-agnostic Explanations) uses

perturbed images to understand how a model is making predictions. Taking in a model and a single image as an input, LIME produces output images that show the parts (known as super-pixels) of the original image that were used to classify it.

**Ethics**

As this research used only fully anonymized, routinely collected data, ethical approval was not required in accordance with the NHS Health Research Authority (HRA) guidelines (http://www.hra-decisiontools.org.uk/ethics/") Therefore written consent from participants was waived.

**Author contribution**

B.P, D.B and J.H contributed to the concept of the study, H.H prepared the data, C.M pre-processed the data, developed the models and analysed the results. B.P, J.H and C.M designed the modelling pipeline. C.M drafted the paper and B.P, J.H, D.B and H.H critically reviewed paper.

**Data availability:**

The data that support the findings of this study are available from the corresponding author upon reasonable request.

The code used to build the models is available at https://github.com/CMapstone/eMLife.

**Competing interests:**

The authors declare no competing interests

**Acknowledgements**

We would like to thank other members (current and past) of the Plusa lab for their discussion about the paper. C.M. was sponsored by the Wellcome Trust as part of the Quantitative and Biophysical Biology programme (220001/Z/19/Z). Funding to B.P Lab was provided by the Wellcome Trust grant Seed Awards in Science (212372/Z/18/Z). D.B was funded by Manchester University NHS Foundation Trust.

## Figures

### Figure 1

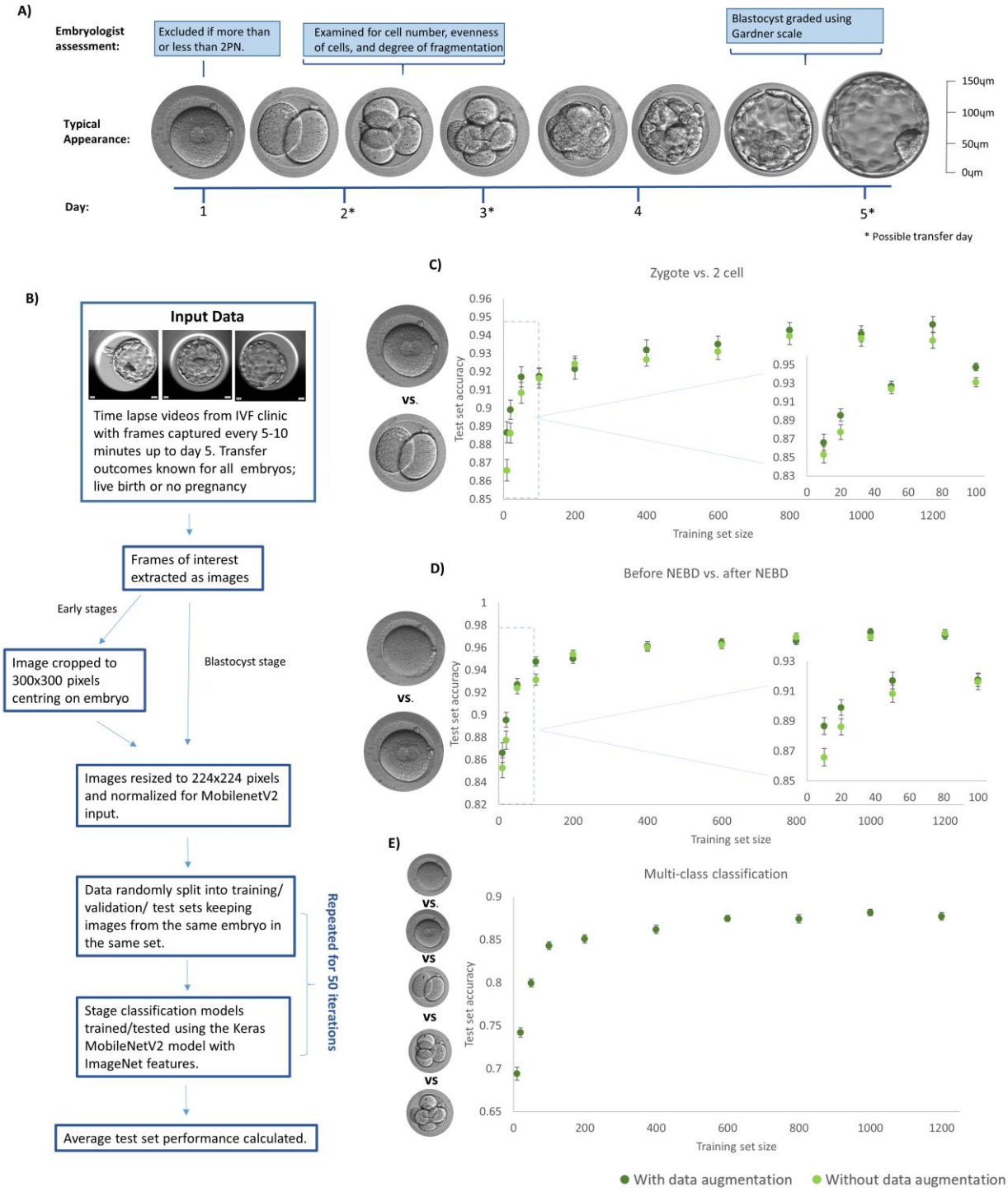



**Figure 1: Performance of developmental stage classification models with increasing amount of training data** A) The typical development of human embryo and an overview of embryologist assessments carried out at each stage. B) The methodology of training developmental stage classification models from time-lapse data. C) The accuracy of the test set on models trained to classify an image as zygote or 2 cell with varying amounts of training data. D) The accuracy of the

test set on models trained to classify an image as before NEBD or after NEBD with varying amounts of training data. E) The accuracy of the test set on multi-class models trained to classify an image into one out of 5 developmental stage classes with varying amounts of training data. C-E) all show the average scores over 50 training attempts with 25 embryos in each class for the test and validation set.

**Figure 2**

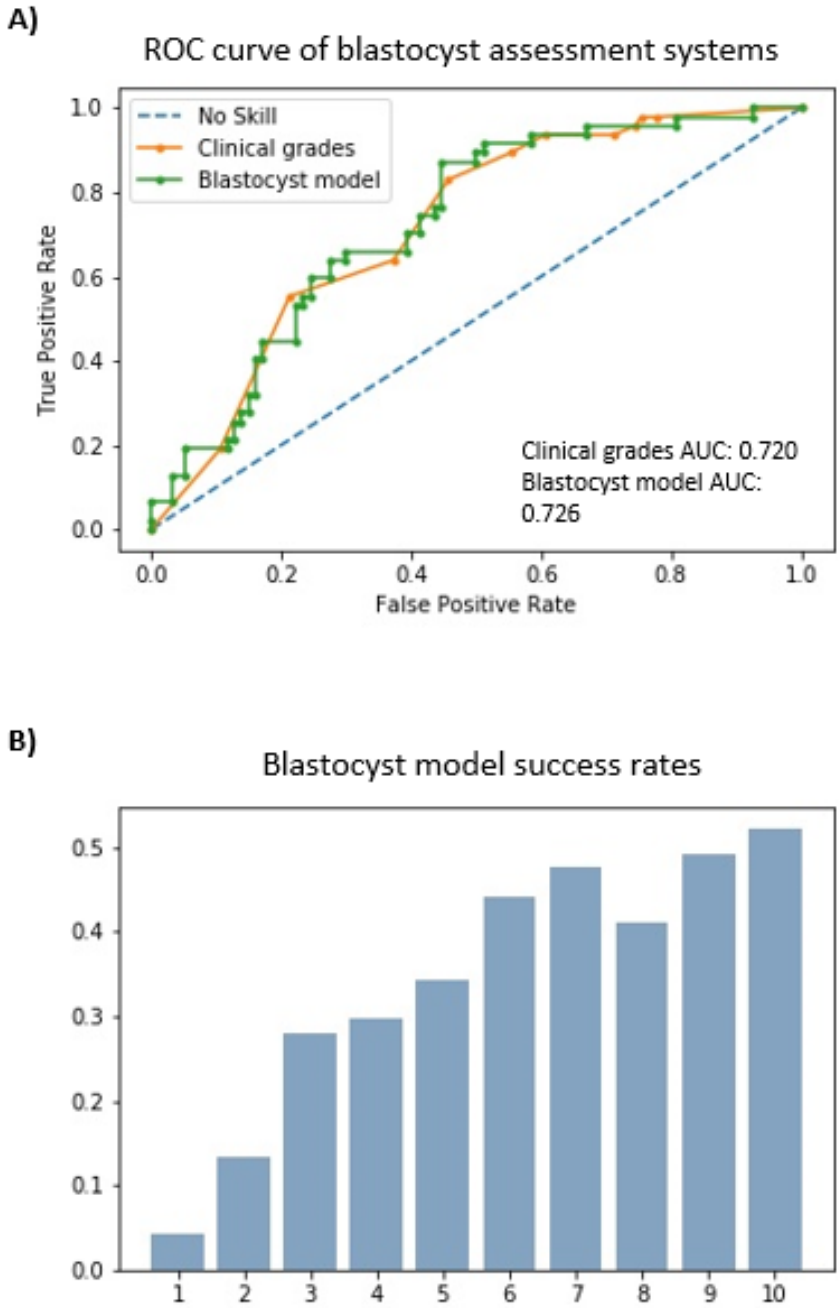



**Figure 2: Live birth predicted from blastocyst stage.** Only embryos for which clinical grades were available were included, a total of 141. Clinical grades were originally based on the Gardner scoring system[28], these were then converted to numerical grades to allow a ROC curve to be generated. The blastocyst model predicts live birth using the last frame of the time-lapse video on day 5, the predictions produced by this model are a number between 0 and 1, 0 corresponding to an

unsuccessful transfer (no pregnancy) and 1 corresponding to a successful transfer (resulting in live birth).

**Figure 3**

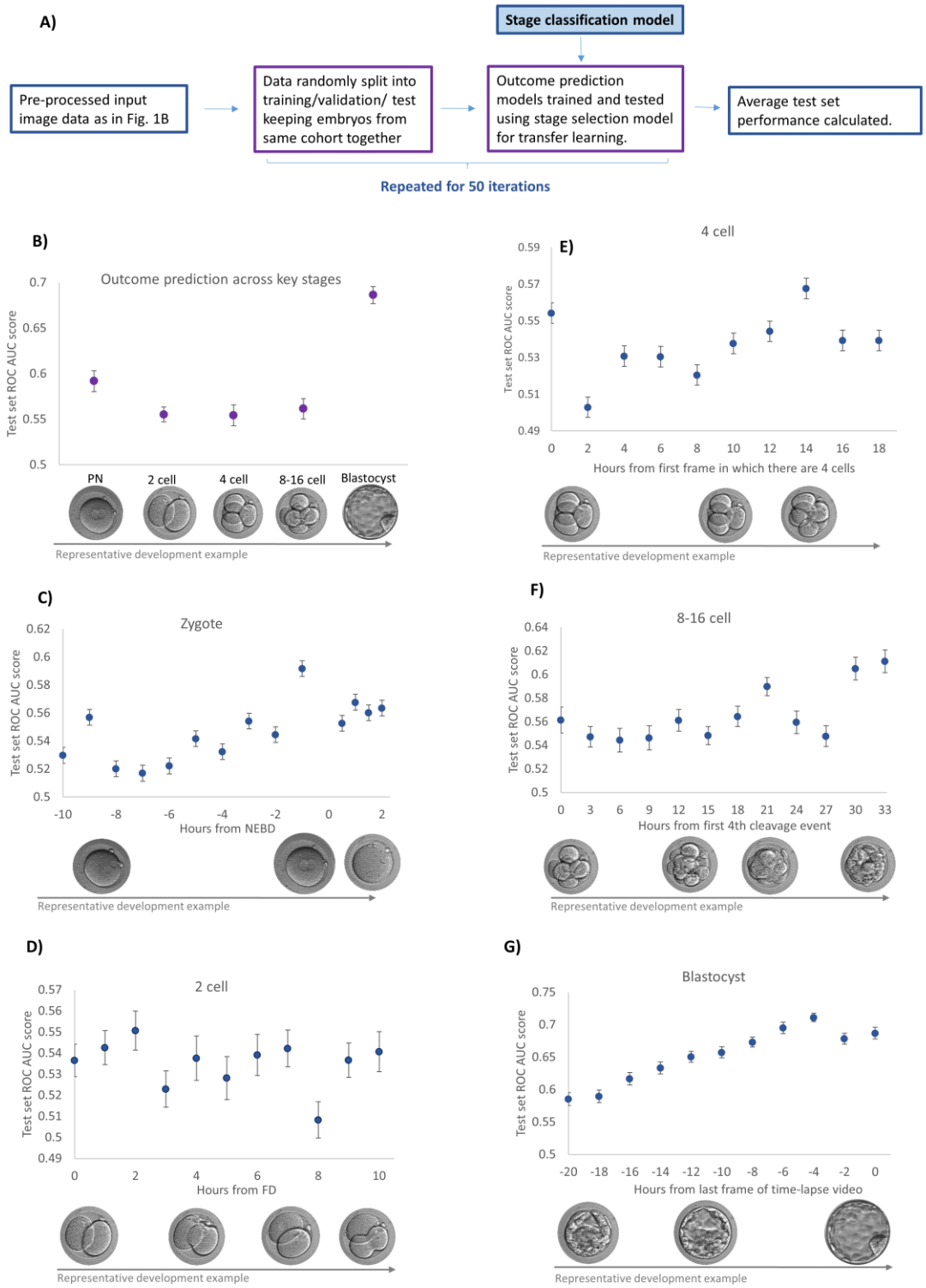
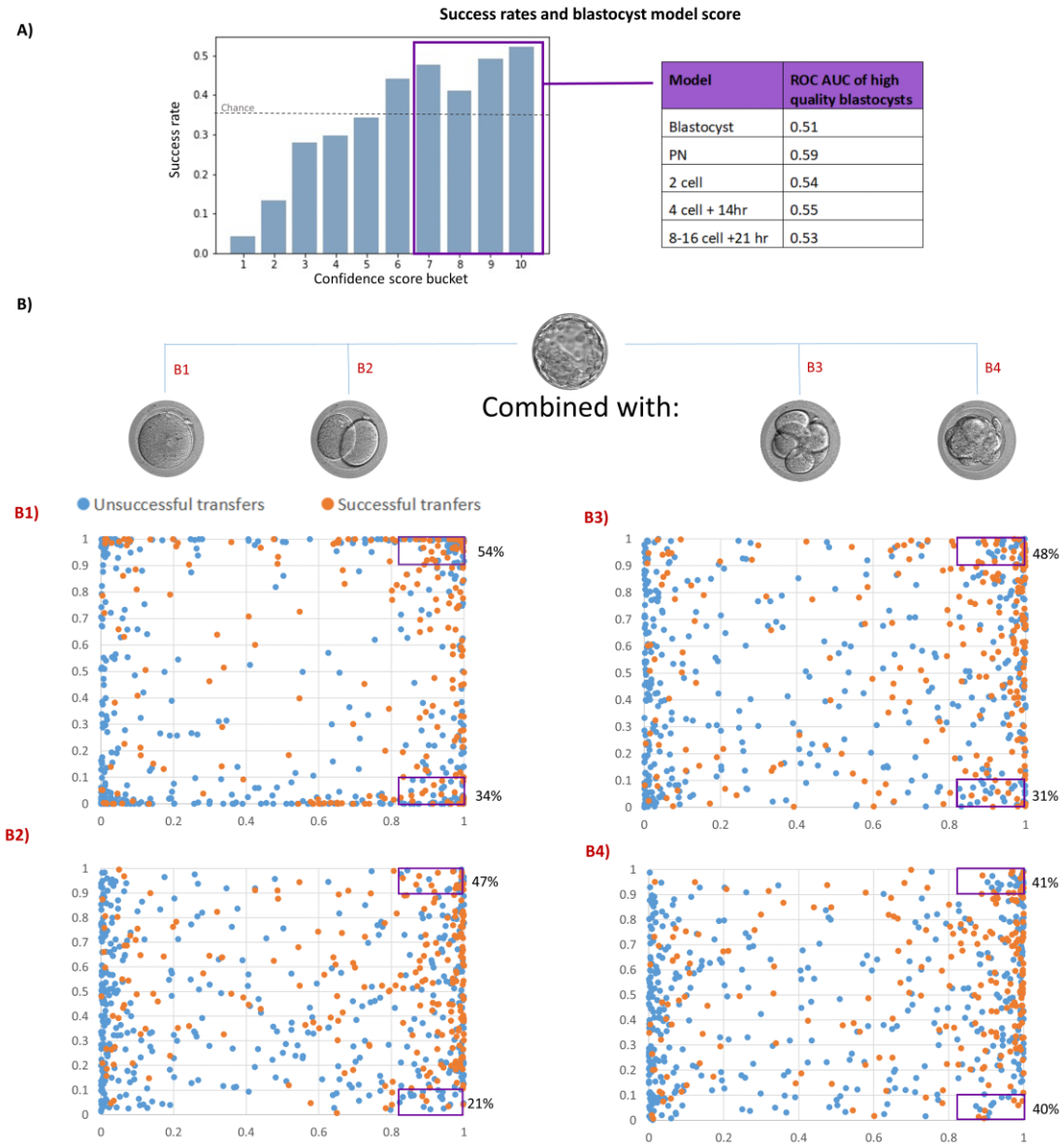



**Figure 3: Predicting Live birth from a transferred embryo using specific time-points across development.** A) The methodology for developing the live birth prediction models B) the ROC AUC

score of the test set for models trained to predict transfer outcome using images from precisely defined moments of development. 'PN 'is one hour before completion of NEBD, '2 cell' is first frame with 2 cells, '4 cell' is first frame with 4 cells, '8-16' cell is the first 4$^{th}$ cleavage event, and 'blastocyst' is the last frame of the video. C) the ROC AUC score of the test set for models trained to predict transfer outcome using images at various time intervals before and after NEBD. D) the ROC AUC score of the test set for models trained to predict transfer outcome using images at various time intervals after FD. E) The ROC AUC score of the test set for models trained to predict transfer outcome using images at various time intervals after the first frame with 4 cells. F) the ROC AUC score of the test set for models trained to predict transfer outcome using images at various time intervals after the first 4$^{th}$ cleavage event. G) the ROC AUC score of the test set for models trained to predict transfer outcome using images at various time intervals before the last frame of the time lapse video. B) – G) all show the average scores over 50 training attempts with 25 embryos in the successful class and 48 embryos in the unsuccessful class for both the test and validation set.

**Figure 4**




| Model | ROC AUC of high quality blastocysts |
|---|---|
| Blastocyst | 0.51 |
| PN | 0.59 |
| 2 cell | 0.54 |
| 4 cell + 14hr | 0.55 |
| 8-16 cell +21 hr | 0.53 |

**Figure 4: Ranking Embryos by blastocyst model score and combining model outputs.** A) The association between live birth success rate and blastocyst model prediction (a model trained to predict live birth from images taken at blastocyst stage). The embryos have been sorted into buckets by blastocyst model confidence score, bucket 1 is lowest and bucket 10 is highest. The graph (left) shows the success rate of each bucket, defined as the fraction of embryos in each bucket that resulted in a live birth. The table on the right shows the ROC AUC for embryos in buckets 7-10 ("high quality blastocysts") obtained using each of the models trained in this study. B1-4) Blastocyst model score vs. PN, 2 cell, 4 cell +14hr, and 8-16 cell+ 21hr model score respectively. The probability of live birth for high quality blastocysts (see A above) with a very high pre-blastocyst model score (>0.9) or very low pre-blastocyst model score (<0.1) is also shown for each graph

**Tables**

| Stage | ROC AUC when standard MobileNetv2 model used | ROC AUC when extra transfer learning used |
|---|---|---|
| PN | 0.547 | 0.557 |
| 2 cell | 0.555 | 0.520 |
| 4 cell | 0.533 | 0.578 |
| 8-16 cell | 0.578 | N/A |
| Blastocyst | 0.714 | 0.703 |

**Table 1:** Performance of a hold out test using either the original MobileNetv2 model with fixed features up to the last layer or the MobileNetv2 model with fixed features and extra transfer learning; an extra hidden layer before the last layer has been pre-trained as a stage classifier. The number of hidden units in the last layer is the amount we found to be optimal for that developmental stage.

| Hour | Percent of embryos cavitating | |
|---|---|---|
| | Successful | Unsuccessful |
| 21 | 7.3% | 4.8% |
| 24 | 27.4% | 21.9% |
| 27 | 50.9% | 40.0% |
| 30 | 81.2% | 50.8% |

**Table 2:** Percent of embryos that have begun cavitating in each group at various time-points after first 4$^{th}$ cleavage event

| Stage | Average Individual ROC AUC | ROC AUC using average confidence score |
|---|---|---|
| PN | 0.591 | 0.603 |
| 2 cell | 0.555 | 0.575 |
| 4 cell | 0.567 | 0.581 |
| 8-16 cell | 0.585 | 0.585 |
| Blastocyst | 0.686 | 0.680 |

**Table 3:** Effect of averaging score from many models rather than using just one model

## References


1 Bhardwaj, A., Kishore, S. & Pandey, D. K. Artificial Intelligence in Biological Sciences. *Life (Basel, Switzerland)* **12**, 1430 (2022). https://doi.org:10.3390/life12091430

2 Bhosale, Y. H. & Patnaik, K. S. Application of Deep Learning Techniques in Diagnosis of Covid-19 (Coronavirus): A Systematic Review. *Neural processing letters*, 1-53 (2022). https://doi.org:10.1007/s11063-022-11023-0

3 Anwar, S. M. *et al.* Medical Image Analysis using Convolutional Neural Networks: A Review. *Journal of medical systems* **42**, 226-213 (2018). https://doi.org:10.1007/s10916-018-1088-1

4 Niakan, K. K., Han, J., Pedersen, R. A., Simon, C. & Pera, R. A. R. Human pre-implantation embryo development. *Development (Cambridge)* **139**, 829-841 (2012). https://doi.org:10.1242/dev.060426

5 Piliszek, A., Grabarek, J. B., Frankenberg, S. R. & Plusa, B. Cell fate in animal and human blastocysts and the determination of viability. *Molecular human reproduction* **22**, 681-690 (2016). https://doi.org:10.1093/molehr/gaw002

6 Płusa, B. & Piliszek, A. Common principles of early mammalian embryo self-organisation. *Development (Cambridge)* **147**, dev183079 (2020). https://doi.org:10.1242/dev.183079

7 Bormann, C. L. *et al.* Consistency and objectivity of automated embryo assessments using deep neural networks. *Fertility and sterility* **113**, 781-787.e781 (2020). https://doi.org:10.1016/j.fertnstert.2019.12.004

8 Kragh, M. F. & Karstoft, H. Embryo selection with artificial intelligence: how to evaluate and compare methods? *Journal of assisted reproduction and genetics* **38**, 1675-1689 (2021). https://doi.org:10.1007/s10815-021-02254-6

9 Khosravi, P. *et al.* Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ digital medicine* **2**, 21-21 (2019). https://doi.org:10.1038/s41746-019-0096-y

10 Kragh, M. F., Rimestad, J., Berntsen, J. & Karstoft, H. Automatic grading of human blastocysts from time-lapse imaging. *Comput Biol Med* **115**, 103494 (2019). https://doi.org:10.1016/j.compbiomed.2019.103494

11 Tran, D., Cooke, S., Illingworth, P. J. & Gardner, D. K. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Human reproduction (Oxford)* **34**, 1011-1018 (2019). https://doi.org:10.1093/humrep/dez064

12 Kanakasabapathy, M. K. *et al.* Development and evaluation of inexpensive automated deep learning-based imaging systems for embryology. *Lab on a chip* **19**, 4139-4145 (2019). https://doi.org:10.1039/c9lc00721k

13 Thirumalaraju, P. *et al.* Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality. *Heliyon* **7**, e06298-e06298 (2021). https://doi.org:10.1016/j.heliyon.2021.e06298

14 Liao, Q. *et al.* Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. *Communications biology* **4**, 415-415 (2021). https://doi.org:10.1038/s42003-021-01937-1

15 Chavez-Badiola, A. *et al.* Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Scientific reports* **10**, 4394-4394 (2020). https://doi.org:10.1038/s41598-020-61357-9

16 Huang, B. *et al.* Using deep learning to predict the outcome of live birth from more than 10,000 embryo data. *BMC pregnancy and childbirth* **22**, 36-36 (2022). https://doi.org:10.1186/s12884-021-04373-5

17 Lee, C. I. *et al.* End-to-end deep learning for recognition of ploidy status using time-lapse videos. *J Assist Reprod Genet* **38**, 1655-1663 (2021). https://doi.org:10.1007/s10815-021-02228-8

18 Fitz, V. W. *et al.* Should there be an "AI" in TEAM? Embryologists selection of high implantation potential embryos improves with the aid of an artificial intelligence algorithm. *J Assist Reprod Genet* **38**, 2663-2670 (2021). https://doi.org:10.1007/s10815-021-02318-7
19 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* **35**, 770-784 (2020). https://doi.org:10.1093/humrep/deaa013
20 Diakiw, S. M. *et al.* Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. *Human reproduction (Oxford)* **37**, 1746-1759 (2022). https://doi.org:10.1093/humrep/deac131
21 Payá, E., Bori, L., Colomer, A., Meseguer, M. & Naranjo, V. Automatic characterization of human embryos at day 4 post-insemination from time-lapse imaging using supervised contrastive learning and inductive transfer learning techniques. *Computer methods and programs in biomedicine* **221**, 106895-106895 (2022). https://doi.org:10.1016/j.cmpb.2022.106895
22 Miyagi, Y., Habara, T., Hirata, R. & Hayashi, N. Feasibility of predicting live birth by combining conventional embryo evaluation with artificial intelligence applied to a blastocyst image in patients classified by age. *Reprod Med Biol* **18**, 344-356 (2019). https://doi.org:10.1002/rmb2.12284
23 Ueno, S. *et al.* Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: a single-center large cohort retrospective study. *Fertil Steril* **116**, 1172-1180 (2021). https://doi.org:10.1016/j.fertnstert.2021.06.001
24 Sawada, Y. *et al.* Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. *Reprod Biomed Online* **43**, 843-852 (2021). https://doi.org:10.1016/j.rbmo.2021.05.002
25 Roberts, S. A., Hirst, W. M., Brison, D. R. & Vail, A. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. *Human reproduction (Oxford)* **25**, 2792-2802 (2010). https://doi.org:10.1093/humrep/deq213
26 Deshpande, P. & Gupta, A. Causes and prevalence of factors causing infertility in a public health facility. *Journal of human reproductive sciences* **12**, 287-293 (2019). https://doi.org:10.4103/jhrs.JHRS_140_18
27 Castillo, C. M. *et al.* The impact of IVF on birthweight from 1991 to 2015:a cross-sectional study. (2019).
28 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**, 1155-1158 (2000). https://doi.org:10.1016/S0015-0282(00)00518-5
29 Balaban, B. *et al.* Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Reproductive biomedicine online* **22**, 632-646 (2011). https://doi.org:10.1016/j.rbmo.2011.02.001
30 Ribeiro, M.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the NAACL-HLT 2016–2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session,* San Diego, CA, USA, 12–17 June 2016.
31 Omer, S. & Lior, R. Ensemble learning: A survey. *WIREs Data Mining Knowl Discov.* **8**, e1249 (2018).
32 Braude, P., Bolton, V. & Moore, S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature (London)* **332**, 459-461 (1988). https://doi.org:10.1038/332459a0

33 Dobson, A. T. *et al.* The unique transcriptome through day 3 of human preimplantation development. *Human molecular genetics* **13**, 1461-1470 (2004). https://doi.org:10.1093/hmg/ddh157

34 Glujovsky, D., Blake, D., Farquhar, C. & Bardach, A. Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. *Cochrane database of systematic reviews* **7**, CD002118-CD002118 (2012). https://doi.org:10.1002/14651858.CD002118.pub4

35 Castillo, C. M. *et al.* The impact of selected embryo culture conditions on ART treatment cycle outcomes: a UK national study. *Human reproduction open* **2020**, hoz031 (2020). https://doi.org:10.1093/hropen/hoz031

36 Zhu, Q. *et al.* Live birth rate and neonatal outcome following cleavage-stage embryo transfer versus blastocyst transfer using the freeze-all strategy. *Reproductive biomedicine online* **38**, 892-900 (2019). https://doi.org:10.1016/j.rbmo.2018.12.034

37 De Vos, A. *et al.* Birthweight of singletons born after cleavage-stage or blastocyst transfer in fresh and warming cycles. *Human reproduction (Oxford)* **33**, 196-201 (2018). https://doi.org:10.1093/humrep/dex361

38 Wang, C. *et al.* Leukocyte telomere length in children born following blastocyst-stage embryo transfer. *Nature medicine* **28**, 2646-2653 (2022). https://doi.org:10.1038/s41591-022-02108-3

39 Brison, D. R. IVF children and healthy aging. *Nature medicine* **28**, 2476-2477 (2022). https://doi.org:10.1038/s41591-022-02098-2

40 Glujovsky, D., Farquhar, C., Quinteiro Retamar, A. M., Alvarez Sedo, C. R. & Blake, D. Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. *Cochrane database of systematic reviews* **2016**, CD002118-CD002118 (2016). https://doi.org:10.1002/14651858.CD002118.pub5

41 Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 4510-4520 (2018).

42 J.Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern recognition*, Miami, USA, pp. 248–255, 2009. DOI: https://doi.org/10.1109/CVPR.2009.5206848.