Fuzzy Systems and Soft Computing
ISSN : 1819-4362

ENHANCEMENTS ON EMBRYO SELECTION WITH ARTIFICIAL INTELLIGENCE

D.VIJAYA SHREE ,Ph.D Research Scholar, SRI RAMAKRISHNA COLLEGE OF ARTS AND SCIENCE

Dr.B.L.SHIVA KUMAR, PRINCIPAL & SECREATARY , SRI RAMAKRISHNA COLLEGE OF ARTS AND SCIENCE

ABSTRACT

A healthy live birth is the aim of an IVF cycle. Even with all of the advancements made in assisted reproductive technology, it is still not possible to anticipate an IVF cycle's outcome with any degree of accuracy. The process used to choose an embryo for transfer is one explanation for this. The conventional approach to determining which embryo to transplant and assessing embryo quality is morphological examination. IVF success rates are subpar as a result of the inter- and intra-observer variability caused by this subjective way of evaluating embryos. It is customary to transfer many embryos in order to get around this, which could lead to high-risk multiple pregnancies.

The results are still not ideal, despite the introduction of pre-implantation genetic testing for aneuploidy and time-lapse incubators to assist raise the likelihood of a live birth. In order to enhance the success of in vitro fertilisation (IVF), artificial intelligence (AI) is being used more and more in the medical industry. Numerous research works have been released that explore the application of AI as an automated, objective method for evaluating embryos. The most current developments in AI for embryology are outlined in this overview. Several methods have been developed recently that use deep learning and artificial intelligence (AI) to enhance and automate the process.Artificial intelligence (AI) algorithms are trained to automatically rate embryos according to their likelihood of a successful implantation based on pictures of embryos with known implantation data (KID). A sizable dataset from 18 IVF centres, totaling 115,832 embryos—14,644 of which were transplanted KID embryos—was used to train and assess the model. It was demonstrated that the completely automated iDAScore v1.0 model outperformed a cutting-edge manual embryo selection approach by at least as much. Additionally, biases resulting from inter- and intraobserver variation are eliminated when embryo grading is fully automated, which implies less manual evaluations.

Keywords: In Vitro Fertilisation (IVF), Artificial Intelligence (AI),Known Implantation Data (KID)

1.INTRODUCTION

Artificial intelligence (AI) has had a significant impact on in vitro fertilisation (IVF) research and innovation during the past few years. AI applications have the potential to support or even completely automate in vitro fertilisation (IVF) processes in the near future. These processes include gamete quality assessment, sperm selection during intracytoplasmic sperm injection (ICSI), oocyte collection, donor matching, patient stimulation protocols, and the selection and ranking of embryos for cryopreservation and transfer [1]. Additionally, by implementing predictive maintenance in IVF equipment and automatically extracting and analysing important performance data to undertake continuous quality control, AI may aid in the optimisation and standardisation of clinical procedures. Identifying the most viable embryo for transfer through embryo assessment has been a difficult task since the beginning of in vitro fertilisation (IVF).Continuous monitoring of embryo growth in vitro has been made possible by the introduction of time-lapse photography into clinical procedures [1]. This has made it possible to identify morphological alterations and events at their precise time of occurrence [2].Numerous models of embryo selection have been devised [3–7] based on these morphokinetic characteristics. Both blastocyst prediction [8, 9], genetic status [10–12], gestational sacs [4], and live birth [13–15] were the endpoints of these models. When time-lapse and morphokinetic selection are utilised instead of normal incubation, investigations have generally demonstrated an improvement [16, 17], while some studies have discovered that internal model validation is necessary before usage[18].

2.RELATED WORK

Many couples and individuals are turning to assisted reproductive treatments in order to help with conception as a result of the general decline in worldwide fertility (GBD, 2018). Sadly, only about

20–30% of IVF attempts result in a pregnancy (Wang and Sauer, 2006), which puts a heavy emotional and financial burden on those trying to conceive. The quality of the embryos produced during the IVF procedure is one of the most important factors in determining the success of the pregnancy, and the embryo selection process is crucial to guaranteeing the patient the quickest possible time to conception. Improving the selection of embryos for uterine transfer during in vitro fertilisation (IVF) is highly motivated.

Currently, choosing embryos is a manual procedure that involves skilled clinical embryologists examining each embryo under an optical light microscope to visually analyse its morphological traits. The Gardner Scale (Gardner and Sakkas, 2003) is the most widely used scoring system among embryologists. It assesses and grades morphological characteristics such inner cell mass (ICM) quality, trophectoderm quality, and embryo developmental advancement using an alphanumeric scale. The considerable degree of subjectivity and intra- and inter-operator variability that exists across embryologists of varying skill levels is one of the main issues in embryo grading (Storr et al., 2017).AI algorithms and their applications in IVF have been evaluated in a number of publications [2–8]. However, as embryo evaluation and selection is the most actively researched application of AI in IVF at the moment—more than ten publications have been published in 2020—we specifically focus on this topic in this work. For over ten years, research has been conducted on automated embryo evaluation utilising machine learning or computer vision based on embryo photos [9, 10]. However, a lot of the articles from the last few years have been more concerned with competitiveness and commercialization than with methodological innovations and technical aspects of AI [11–15]. Rather, it appears that they are more concerned in presenting big datasets, high performance values derived from several metrics, and the capacity to outperform human and embryologist performance. Studies differ greatly in their clinical objectives and evaluation techniques, and occasionally performance comparisons are based on entirely distinct data foundations (e.g., patient demographics, unbalanced data, sub-cohorts, etc.). It has therefore become obvious that the research community. The goal of the current clinical study was to create and evaluate an assessment that uses non-invasive artificial intelligence (AI).

Furthermore, this strategy is too expensive for many laboratories and clinics due to the need for specialised time-lapse imaging apparatus, which further restricts the technique's widespread application. Using single static two-dimensional pictures obtained by optical light microscopy techniques, the current clinical experiment aimed to design and test a non-invasive artificial intelligence (AI)-based assessment strategy to support embryo selection during IVF. In order to develop a reliable model for the study of Day 5 embryos (blastocysts) and the prediction of clinical pregnancy outcomes, computer vision image processing techniques and deep learning were used.

3.METHODOLOGY

The embryo population and the end result used for training and, more crucially, evaluation are critical information to have when comparing AI models. The AI models utilised for embryo assessment are characterised in this paper using a population-outcome scheme based on their data foundation. The four distinct attributes that comprise the system are depicted in Figure 1.

- Fertilization: Which procedure or methods of fertilisation were used? ICSI, IVF, or both?
- Culture: What was the duration of the embryos' incubation? (For instance, five days)
- Sub-cohort : Which embryos from the available pool were included in the sub-cohort? (For example, euploid, fresh, cryopreserved blastocysts, hatched blastocysts, etc.)
- Result: What was the endpoint that was measured? For instance, a live birth or a foetal heartbeat

While the outcome specifies the clinical goal (or ground truth) that the model was tested against, the embryo population is defined by the first three variables and describes which embryos were included in a study. For example, the ICSI-D5/D6-* embryo population and live birth outcome would be reported in a study predicting live birth on all embryos transferred on day 5 or 6 after fertilisation by ICSI. Only transplanted embryos were taken into consideration by using live birth as the outcome.

Fig. 1 Example scheme for reporting embryo population and outcome. A study reporting prediction of live birth on transferred day 5 blastocysts fertilized by ICSI would have the embryo population *ICSI-D5-Blastocyst-Transfer* and outcome *live birth*

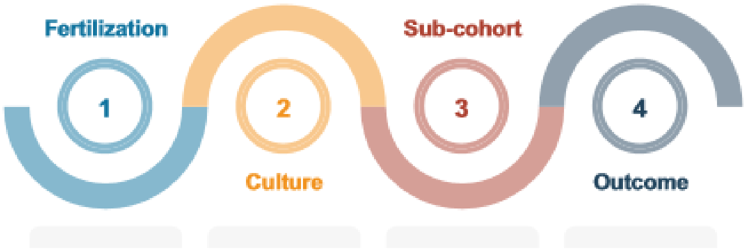
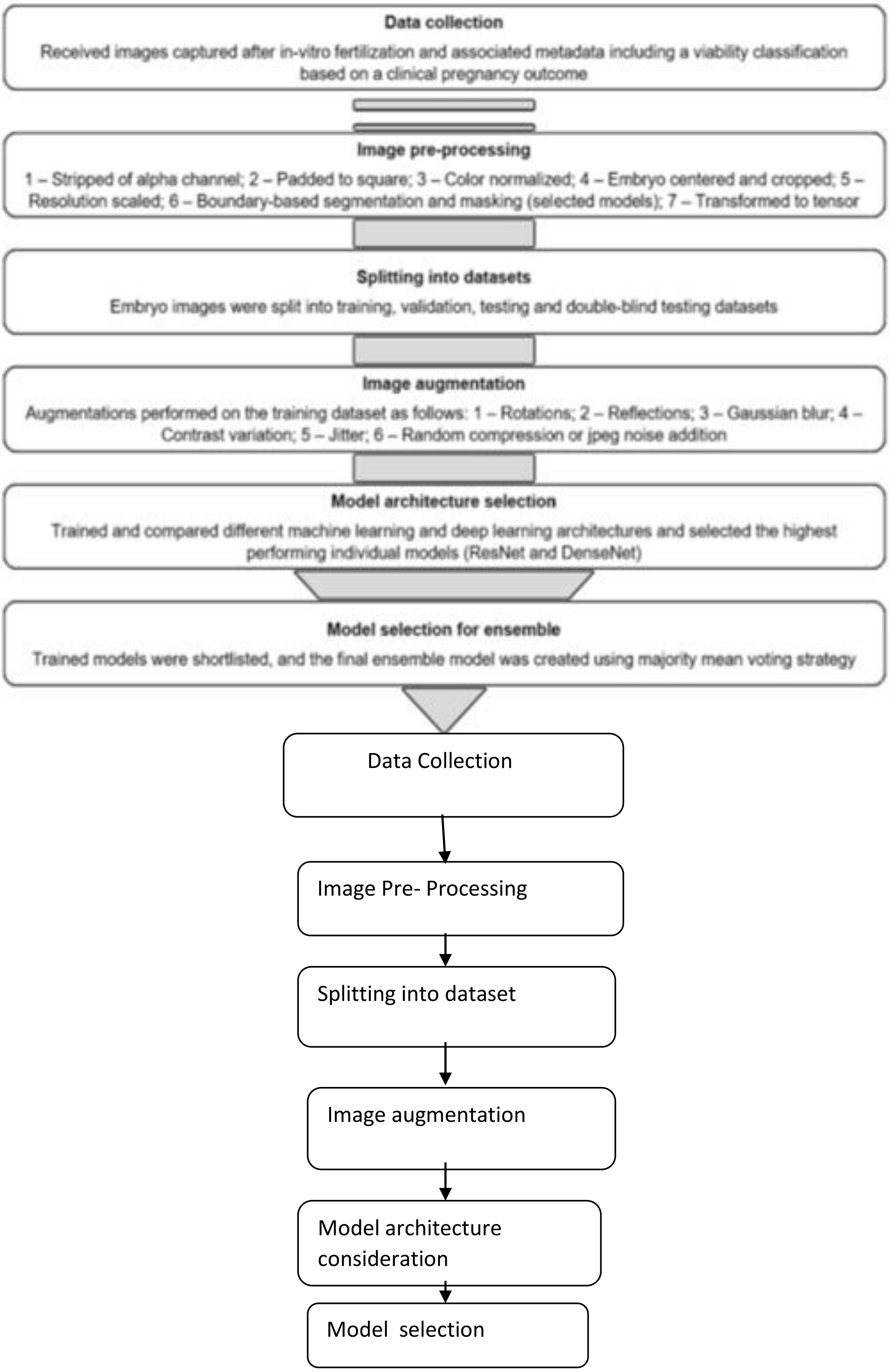


Figure 1: example scheme for reporting embryo population and outcome.

An embryo viability score of 50% or more was deemed viable for the AI model, whereas a score of less than 50% was deemed non-viable. At the moment the picture was taken, Day 5 blastocysts' scores were reported by the embryologist. These results were grouped into "likely viable" and "likely non-viable" groups, which served as the basis for the scoring bands. This generalisation made it possible to compare the AI model's (viable/non-viable) predictions with the binary forecasts made by the embryologists. The Gardner scale of morphokinetic grading (Gardner and Sakkas, 2003) served as the basis for the embryologists' scoring system. Each letter (A–E) represented the grade of the embryo's trophectoderm and ICM.A description of the embryo's developmental stage leading up to hatching or a numerical score was also provided. Following the sequence of increasing embryonic development, numbers were assigned: 1. Denotes the beginning of cavitation; 2. Early blastocyst; 3. Full blastocyst; and 4. Expanded blastocyst5 is the blastocyst hatching. An early blastocyst (>2) was thought to be the embryo's developmental stage if none was specified. By assigning a numerical score between 1 and 5 to the embryologist's evaluation and, correspondingly, splitting the AI model's inferences into five equal bands labelled 1 to 5 (from the least inference to the maximum inference), comparisons of embryo viability ranking were made. An image of an embryo was considered to be in "concordance" if both the AI model and the embryologist assigned the same rating to it.However, this result was marked as "model correct" if the AI model produced a higher rank than the embryologist and the ground-truth outcome was recorded as viable, or if the AI model produced a lower rank than the embryologist and the ground-truth outcome was documented as ineffective. Similarly, an outcome was marked as "embryologist correct" if the AI model produced a lower rank than the embryologist and the ground-truth outcome was reported as viable, or if the AI model produced a higher rank and the outcome was recorded as non-viable.

Image processing techniques for computer vision .The pre-processing step was applied to all image data, as described below. These techniques for computer vision image processing were applied during the model-building process and included into the finished AI model.

To guarantee that every image was encoded in a 3-channel format (such as RGB), the alpha channel was removed from each one. This process preserved the image's visual integrity while removing extra transparency map-related information. These areas of the picture were not utilised. – Every image was resized to square proportions, so that each side matched the longest side of the source file. In addition to guaranteeing that no important parts of the image were cropped, this procedure made sure that the image dimensions were uniform, comparable, and compatible with deep learning techniques, which specifically demand square dimension images as input. – By taking the mean of each RGB channel and dividing it by its mean value, each image's RGB colour was normalised. Then, each channel was multiplied by a constant 100/255 number to guarantee that each image's mean value in RGB space equaled (100, 100, 100). This stage made sure that each image's brightness was normalised and that colour biases were suppressed. After then, each picture was cropped so that the embryo's centre was in the centre of each one. The process involved determining the optimal ellipse function from an elliptical Hough transform, which was computed using the image's binary threshold map. The technique works by first identifying the hard border of the embryo in the image, then cropping the new image's square border so that the new vision's width and height encircle the ellipse's largest radius, and the ellipse's centre becomes the new image's centre. Then, before training, each image was scaled to a lower resolution. – Boundary-based segmentation was one of the numerous pre-processing steps that images went through in order to train certain models. This procedure works by isolating the region of interest—that is, the embryo—from the background of the image and enabling masking, which focuses the model on categorising the embryo's overall morphological shape.Lastly, since deep learning models require this data structure, each image was converted to a tensor rather than a visually displayable image. Standard pre-trained Image Netvalues, mean (0.485,0.456,0.406) and standard deviation (0.299, 0.224, 0.225) were used to get Tensor Normalisation. An example embryo image processed through the first six pre-processing processes mentioned above is shown in Figure 1.


Data collection
Received images captured after in-vitro fertilization and associated metadata including a viability classification based on a clinical pregnancy outcome
Image pre-processing
1 – Stripped of alpha channel; 2 – Padded to square; 3 – Color normalized; 4 – Embryo centered and cropped; 5 – Resolution scaled; 6 – Boundary-based segmentation and masking (selected models); 7 – Transformed to tensor
Splitting into datasets
Embryo images were split into training, validation, testing and double-blind testing datasets
Image augmentation
Augmentations performed on the training dataset as follows: 1 – Rotations; 2 – Reflections; 3 – Gaussian blur; 4 – Contrast variation; 5 – Jitter; 6 – Random compression or jpeg noise addition
Model architecture selection
Trained and compared different machine learning and deep learning architectures and selected the highest performing individual models (ResNet and DenseNet)
Model selection for ensemble
Trained models were shortlisted, and the final ensemble model was created using majority mean voting strategy
Data Collection
Image Pre- Processing
Splitting into dataset
Image augmentation
Model architecture consideration
Model selection

Flow chart for model creation and selection methodology.

The methodology starts from data collection. Each step summarizes the components task used for development of the final AI model. After image Processing and segmentation ,the image were split into datasets and the training dataset prepared by image augmentation. The highest performing individual models were considered

4.MODEL ARCHITECTURES CONSIDERED

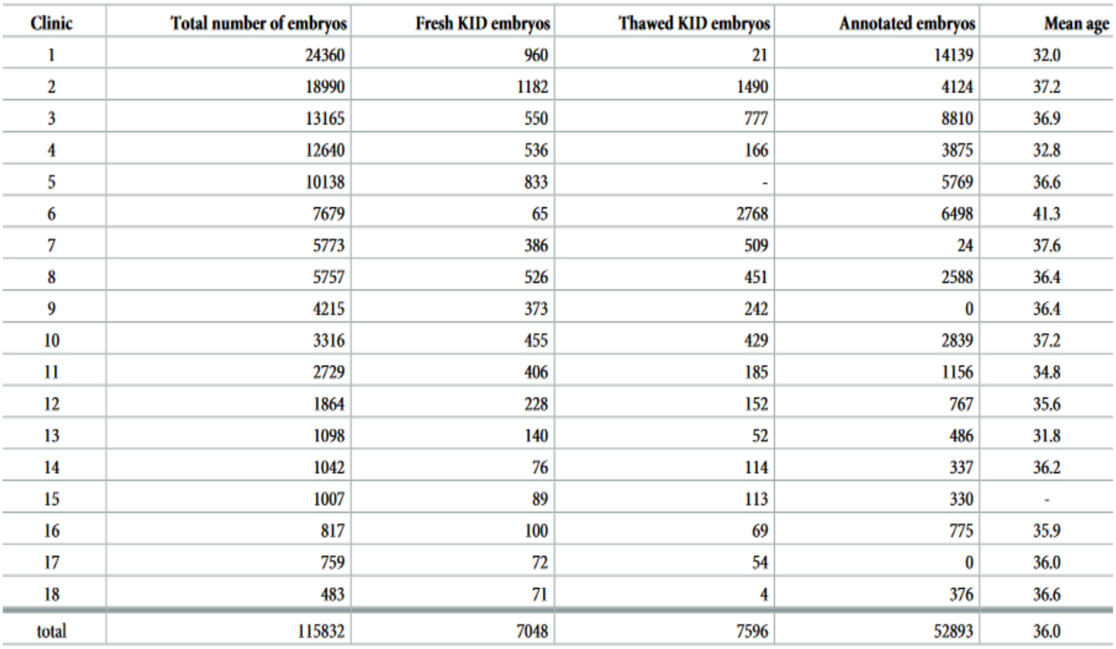
In order to train the AI model, a variety of deep learning and computer vision/machine learning techniques were assessed. Deep learning designs including ResNet-18, ResNet-50, and ResNet-101 (Heetal., 2016) and highly linked networks like DenseNet-121 and DenseNet-161 (Huang et al., 2017) were found to have the greatest impact on the classification of embryo viability. When evaluated separately, these architectures proved to be more reliable than other kinds of models. Other deep learning architectures, such as Inception-ResNetV2 and Inception-V4 (Szegedy et al., 2016), were also evaluated; however, because of their lower individual performance, they were not included in the final AI model.Additionally assessed were computer vision/machine learning models that combined the computation and extraction of computer vision features, such as random forests (Breiman, 2001) and support vector machines (Hearst, 1998). But when tested separately, these techniques produced lower accuracy and less translatability than deep learning techniques, and as a result, they were left out of the final AI model ensemble. See the section on the model selection procedure for further details.

5.DESCRIPTION OF DATA

The investigation includes retrospective data from 18 clinics worldwide from 2011 to 2019. The clinics were chosen without the use of any particular techniques. The total number of embryos, the number of transplanted embryos, and the average age of the females for each clinic are listed in Table 1. Every piece of data used in the research was de-identified and retrospective.

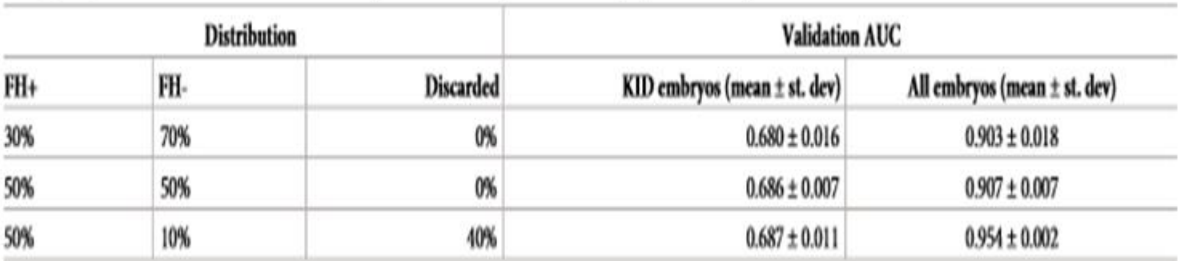
The Act on Research Ethics Review of Health Research Projects in Denmark (Consolidation Act No. 1338 of September 1, 2020) determined that the study in question was not required to be reported to the National Committee on Health Research Ethics.

Table 1. Distribution of total number of embryos, fresh embryos and thawed embryos with known implantation data (KID), total number of annotated embryos and average female age.



Clinic	Total number of embryos	Fresh KID embryos	Thawed KID embryos	Annotated embryos	Mean age
1	24360	960	21	14139	32.0
2	18990	1182	1490	4124	37.2
3	13165	550	777	8810	36.9
4	12640	536	166	3875	32.8
5	10138	833	-	5769	36.6
6	7679	65	2768	6498	41.3
7	5773	386	509	24	37.6
8	5757	526	451	2588	36.4
9	4215	373	242	0	36.4
10	3316	455	429	2839	37.2
11	2729	406	185	1156	34.8
12	1864	228	152	767	35.6
13	1098	140	52	486	31.8
14	1042	76	114	337	36.2
15	1007	89	113	330	-
16	817	100	69	775	35.9
17	759	72	54	0	36.0
18	483	71	4	376	36.6
total	115832	7048	7596	52893	36.0

Table 2. Five-fold cross-validation experiment on varying the distribution between FH+, FH- and discarded embryos used during training. The area under the curve (AUC) was evaluated for both transferred embryos with known implantation data (KID) and all embryos for each validation data set.



Distribution			Validation AUC	
FH+	FH-	Discarded	KID embryos (mean ± st. dev)	All embryos (mean ± st. dev)
30%	70%	0%	0.680 ± 0.016	0.903 ± 0.018
50%	50%	0%	0.686 ± 0.007	0.907 ± 0.007
50%	10%	40%	0.687 ± 0.011	0.954 ± 0.002

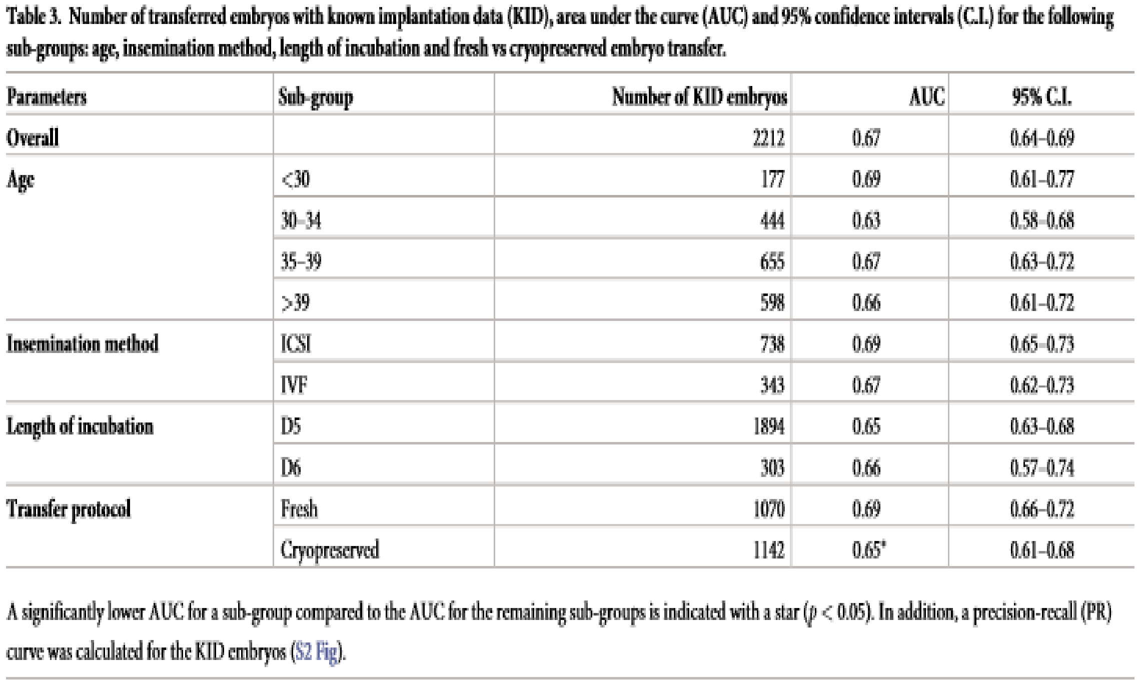


Table 3. Number of transferred embryos with known implantation data (KID), area under the curve (AUC) and 95% confidence intervals (C.I.) for the following sub-groups: age, insemination method, length of incubation and fresh vs cryopreserved embryo transfer.

Parameters	Sub-group	Number of KID embryos	AUC	95% C.I.
Overall		2212	0.67	0.64–0.69
Age	<30	177	0.69	0.61–0.77
	30–34	444	0.63	0.58–0.68
	35–39	655	0.67	0.63–0.72
	>39	598	0.66	0.61–0.72
Insemination method	ICSI	738	0.69	0.65–0.73
	IVF	343	0.67	0.62–0.73
Length of incubation	D5	1894	0.65	0.63–0.68
	D6	303	0.66	0.57–0.74
Transfer protocol	Fresh	1070	0.69	0.66–0.72
	Cryopreserved	1142	0.65*	0.61–0.68

A significantly lower AUC for a sub-group compared to the AUC for the remaining sub-groups is indicated with a star ($p < 0.05$). In addition, a precision-recall (PR) curve was calculated for the KID embryos (S2 Fig).

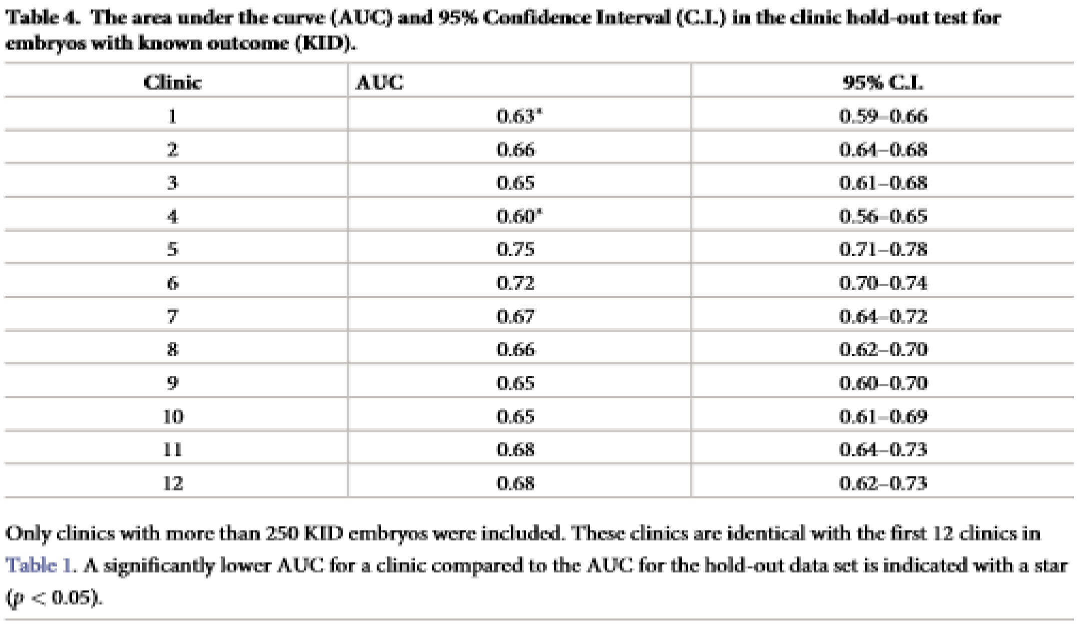


Table 4. The area under the curve (AUC) and 95% Confidence Interval (C.I.) in the clinic hold-out test for embryos with known outcome (KID).

Clinic	AUC	95% C.I.
1	0.63*	0.59–0.66
2	0.66	0.64–0.68
3	0.65	0.61–0.68
4	0.60*	0.56–0.65
5	0.75	0.71–0.78
6	0.72	0.70–0.74
7	0.67	0.64–0.72
8	0.66	0.62–0.70
9	0.65	0.60–0.70
10	0.65	0.61–0.69
11	0.68	0.64–0.73
12	0.68	0.62–0.73

Only clinics with more than 250 KID embryos were included. These clinics are identical with the first 12 clinics in Table 1. A significantly lower AUC for a clinic compared to the AUC for the hold-out data set is indicated with a star ($p < 0.05$).

6.RESULTS

6.1 INITIAL EXPERIMENTS

Three 5-fold cross validation experiments were conducted in order to examine the impact of the embryo sampling strategy used in the training data set. The model was trained only on transplanted KID embryos in the first two studies. These embryos were either oversampled with positive samples (50 percent FH+ and 50 percent FH-) or sampled based on actual prevalence (30% FH+ and 70% FH-). In the third experiment, the model was trained using oversampled positive data (50 percent FH+, 10 percent FH-, and 40 percent discarded) and discarded embryos pseudo-labeled as FH-. For each of the five folds, the performance was assessed on transplanted KID embryos in the validation data set. Table 2 shows that the mean AUC of the three distinct sampling procedures did not change significantly (p > 0.05).

Evaluations were also made on each embryo's overall performance. Compared to the other two sample procedures, the strategy that included discarded embryos had a considerably higher mean AUC (p < 0.005). When discarded embryos were excluded from the two techniques, there was no discernible difference in overall performance. For each of the three techniques, there were distinct patterns in the score distribution of the FH+, FH-, and rejected embryos (S1 Fig). The score distribution of abandoned embryos overlapped with the scores of transplanted FH- and FH+ embryos for models that were exclusively trained on KID embryos. Conversely, there was less overlap between the transplanted embryos and the discarded embryos for models trained on both KID and discarded embryos. The following tests were conducted using a combination of KID and abandoned embryos, with a sampling consisting of 50% FH+, 10% FH-, and 40% discarded embryos, based on the results of these cross-validation trials.\

6.2 FINAL MODEL

The independent test data set, which was not used for training, was used for all ensuing analyses. In total, 17,249 embryos were included in this data collection, of which 2,212 were transferred embryos with known outcomes (KID embryos). Using AUC, the sorting performance of the entire model was assessed. Table 3 shows that the AUC for KID embryos was 0.67, with a 95% confidence interval of 0.64–0.69. The AUC was 0.95 with a 95% confidence interval of 0.95–0.96 if the entire cohort was taken into account.

6.3 SUB-GROUP ANALYSIS

The patient age (for the groups <30, 30–34, 35–39, and >39 years), insemination technique (IVF or ICSI), incubation period (5 or 6 days), and fresh versus cryopreserved embryo transfer were the subgroups that were studied (Table 3). The AUC for KID embryos varied by age group, with the age group of 30-34 years having the lowest AUC, ranging from 0.63 to 0.69. The AUC for ICSI and IVF, in terms of the method of insemination, was 0.69 and 0.67, respectively. For D5 and D6, the AUC was 0.65 and 0.66, respectively, for the duration of incubation. Compared to fresh transfers, which had an AUC of 0.69, cryopreserved embryos had a substantially lower AUC of 0.65.

6.4 CLINIC HOLD-OUT TEST

A clinic hold-out test was conducted to look at how the selected model architecture and training data generalise to new clinics. Table 4 shows that the AUCs for each clinic ranged from 0.60 to 0.75. The AUC was substantially lower for clinics 1 and 4 than it was for the remainingclinics in the test data.

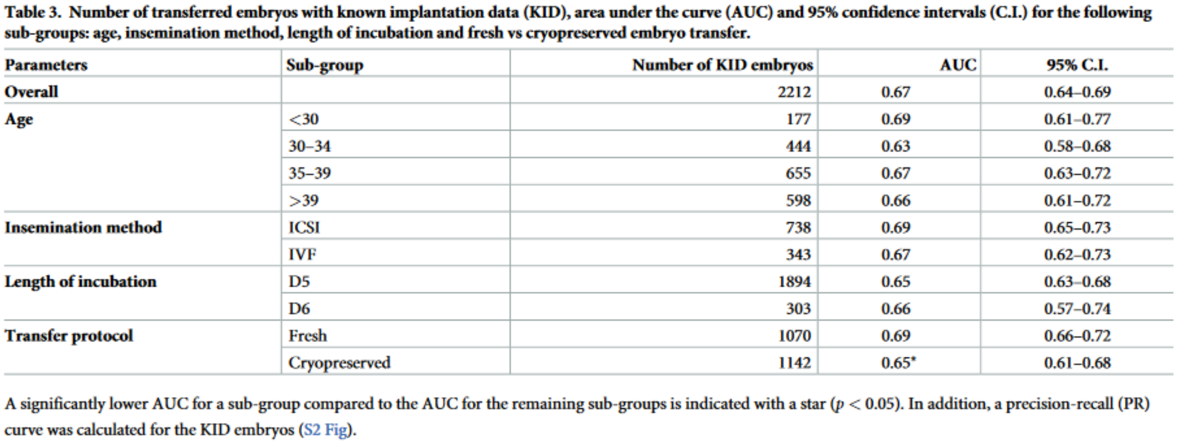


Table 3. Number of transferred embryos with known implantation data (KID), area under the curve (AUC) and 95% confidence intervals (C.I.) for the following sub-groups: age, insemination method, length of incubation and fresh vs cryopreserved embryo transfer.

Parameters	Sub-group	Number of KID embryos	AUC	95% C.I.
Overall		2212	0.67	0.64–0.69
Age	<30	177	0.69	0.61–0.77
	30–34	444	0.63	0.58–0.68
	35–39	655	0.67	0.63–0.72
	>39	598	0.66	0.61–0.72
Insemination method	ICSI	738	0.69	0.65–0.73
	IVF	343	0.67	0.62–0.73
Length of incubation	D5	1894	0.65	0.63–0.68
	D6	303	0.66	0.57–0.74
Transfer protocol	Fresh	1070	0.69	0.66–0.72
	Cryopreserved	1142	0.65*	0.61–0.68

A significantly lower AUC for a sub-group compared to the AUC for the remaining sub-groups is indicated with a star ($p < 0.05$). In addition, a precision-recall (PR) curve was calculated for the KID embryos (S2 Fig).

7.Discussion

7.1 MODEL TRAINING

A primary issue that arises when developing models for embryo selection using implantation data is that a small portion of the entire cohort is assigned a known result. Since the fate of the remaining embryos is unknown, they are not marked. Additionally, because negative labels predominate over positive labels, the tagged embryos are frequently imbalanced. Potential biases may result from these characteristics of the training data and the training techniques[49]. To evaluate three distinct sample strategies for handling unbalanced data and missing labels, we have studied these biases. Specifically, we have trained on KID embryos without any oversampling, on KID embryos with FH+ oversampling, and on both KID embryos and pseudo-labelled discarded embryos.When tested on KID embryos, the results demonstrated that the inclusion of rejected embryos had no impact on performance. Nevertheless, it was evident from an evaluation of all embryos that there was a significant improvement in the ability to identify inferior embryos (Table 2 and S1 Fig). Therefore, generally speaking, using rejected embryos in a model training will provide a completely automated assessment that is relevant to every embryo in a cohort. On the other hand, if models are trained solely on KID embryos, the user has to identify transferable embryos beforehand in order to utilise the model.

7.2 FINAL MODEL

The Compared to the IVY model's 0.93 AUC, the sorting ability for the entire cohort had an AUC of 0.95 [35]. It should be mentioned that every clinic in the IVY model provided data for this study. The total number of embryos increased from 8,836 to 115,832, the number of embryos with known implantation data (KID) increased from 1,773 to 14,644, and the number of FH+ embryos increased from 694 to 4,337 due to the addition of additional new data from these clinics and five additional clinics in this investigation.

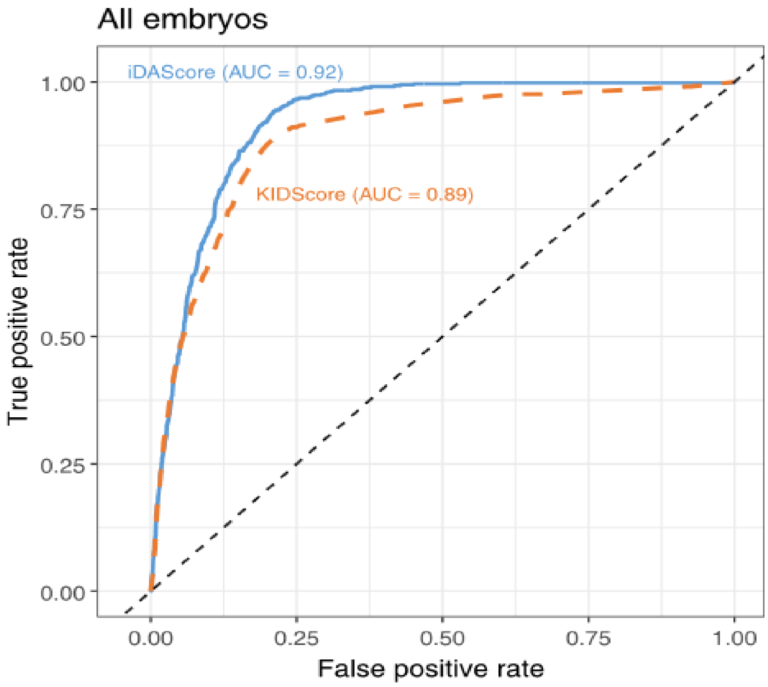


Figure 3: graph for all embryos

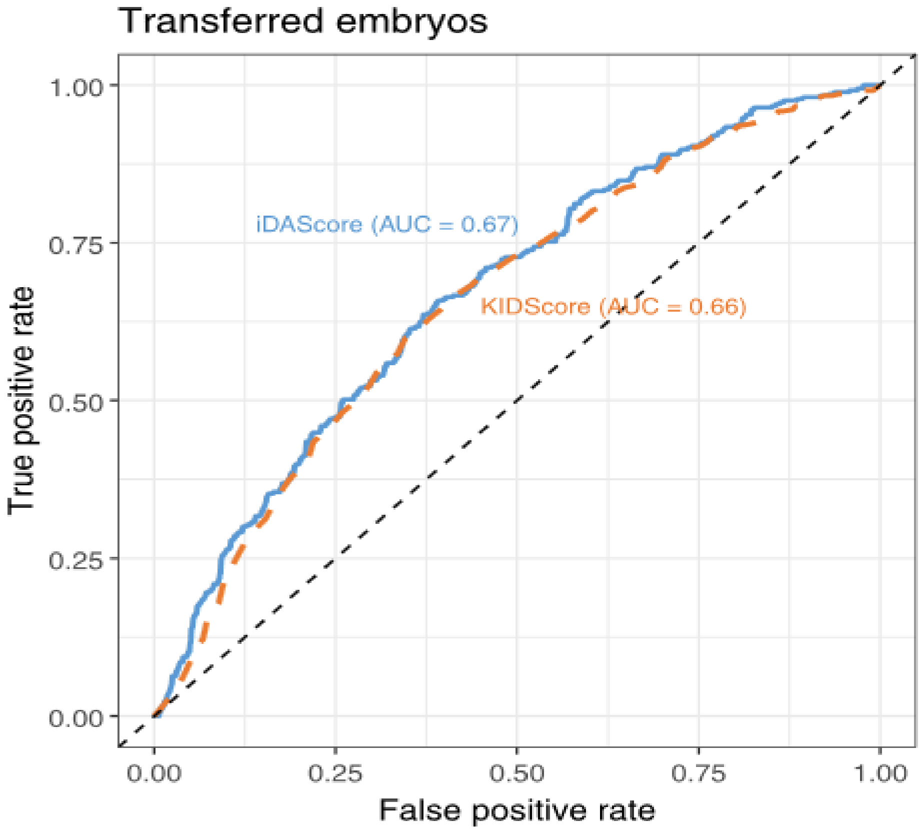


figure 4: graph for transferred embryos

Fig 3. Comparison between ROC curves based on iDAScore (solid blue) and KIDScore (dotted orange). The upper plot shows the ROC curve for the whole embryo cohort (n = 7,932), and the lower plot shows the ROC curve for KID embryos (n = 1,094).

Therefore, even better performance and resilience are achieved because the present data set was more than six times larger than the data set used to train the IVY model. As far as we are aware, this is the biggest data collection available for use in creating an embryo selection model.

Compared to the sorting ability for the entire cohort, which had an AUC of 0.95, the model's sorting abilities for KID embryos had an AUC of 0.67. This discrepancy is predicted since it is far more challenging to separate good morphology blastocysts from the rest of the cohort, which consists of a spectrum of embryos that have been arrested from good morphology blastocysts. The purpose of the model determines which measure is most pertinent. The high AUC for sorting the entire cohort is the most important metric if the goal is to have a completely automated algorithm that sorts between all embryos. The lower AUC for sorting solely within the KID embryos is the most significant, though, if the user prescreens the embryos to identify possible transfer candidates. We suggest that both metrics be included in future publications since they are pertinent to embryo selection algorithms and because their actual clinical environment will determine their practical application.

8 SUB-GROUP ANALYSIS

It is crucial to look at any potential bias within the data set's sub-groups when developing a selection model that will be applied to a variety of clinical settings, patients, and cultural circumstances. The AUC of the overall ROC curve was within the 95% confidence interval for almost all subgroups. The only sub-group with a considerably lower AUC than the others was the cryopreserved sub-group. This is most likely because a successful implantation is dependent on other processes (such as the vitrification/warming and subsequent endometrial preparation). As a result, the sorting gets harder, which lowers the AUC. As far as we are aware, no alternative selection methods have been tried with various subsets in a sizable, independent data set. It is crucial to verify that models perform similarly across various subsets and are free from general biases. It becomes much more crucial to test on age sub-groups when an age parameter is used in a selection model. Given that age is one of the best indicators of a successful implantation, its inclusion will undoubtedly greatly enhance the overall performance of the model. Age may not, however, enhance the sorting ability at the treatment level (that is, the embryo cohort of a single patient), which is essentially what an embryo selection model is all about. Age is a factor that is input into the models of both the STORK algorithm [28] and the AIR E [32]. But since no analysis of age subgroups was done, it is still unknown how well these models can categorise patients based on their course of treatment. A subset of the training data is used to test the model as part of the sub-group analysis, which is an internal validation process [50]. This could provide a positive performance estimate as the testConsidering potential bias among the sub-groups of the data set is essential for creating a selection model that will be used with a range of patients, clinical settings, and cultural contexts. For nearly all subgroups, the total ROC curve's AUC fell inside the 95% confidence interval. The cryopreserved sub-group was the only one with a significantly lower AUC than the others. This is most likely due to the fact that other procedures—like the vitrification/warming and subsequent endometrial preparation—are necessary for a successful implantation. Consequently, the sorting becomes more difficult, lowering the AUC. To the best of our knowledge, no other selection techniques have been used to different subsets within a large, independent data set. addressed using an external validation, which is even more robust and involves testing the model under novel circumstances like new clinics, time periods, procedures, or populations that weren't employed during model development [50]. Another method of internal validation is the geographic validation, often known as the clinic hold-out test.

Hold-out test in the clinic Prior to clinical application, a number of studies have demonstrated the necessity of evaluating selection models on internal data [18, 51]. We employed a clinical hold-out validation technique in this investigation, where the model was tested on a particular clinic after it had been trained on data excluding it. Table 4 illustrates how these models sorted similarly, with most falling within the 95% confidence interval of the However, as no age sub-group analysis was done, it is still unknown how well these models sorted at the treatment level. A subset of the training data is used to test the model as part of the sub-group analysis, which is an internal validation process [50]. This could provide a positive performance estimate as the testConsidering potential bias among the sub-groups of the data set is essential for creating a selection model that will be used with a range of patients, clinical settings, and cultural contexts. For nearly all subgroups, the total ROC curve's AUC fell inside the 95% confidence interval. The cryopreserved sub-group was the only one with a significantly lower AUC than the others. This is most likely due to the fact that other procedures—like the vitrification/warming and subsequent endometrial preparation—are necessary for a successful implantation. Consequently, the sorting becomes more difficult, lowering the AUC. To the best of our knowledge, no other selection techniques have been used to different subsets within a large, independent data set. addressed using an external validation, which is even more robust and involves testing the model under novel circumstances like new clinics, time periods, procedures, or populations that weren't employed during model development [50]. Another method of internal validation is the geographic validation, often known as the clinic hold-out test.ultimate model. The AUC was much lower in Clinics 1 and 4 with young women (Table 1) than it was on the test set. Most of the transferred KID embryos for the younger women in these clinics were likely very good blastocysts. A more diversified cohort of KID embryos is probably transferred for the older mothers in the other facilities. The AUCs were lower for clinics 1 and 4 because sorting within a homogeneous top-quality blastocysts is more challenging than sorting within a more heterogeneous collection of KID embryos. Therefore, the variation in AUC is most likely due to a bias in the evaluation of pre-selected embryos for transfer rather than a problem with the model's performance. Note that a high implantation rate does not always correspond with a high AUC, and vice versa. Specifically, a high AUC indicates how well a model classifies embryos within a given cohort. As a result, variations in AUCs may be caused by inadequate model generalisation, as well as by variations in the patient population and clinical procedures mentioned previously.

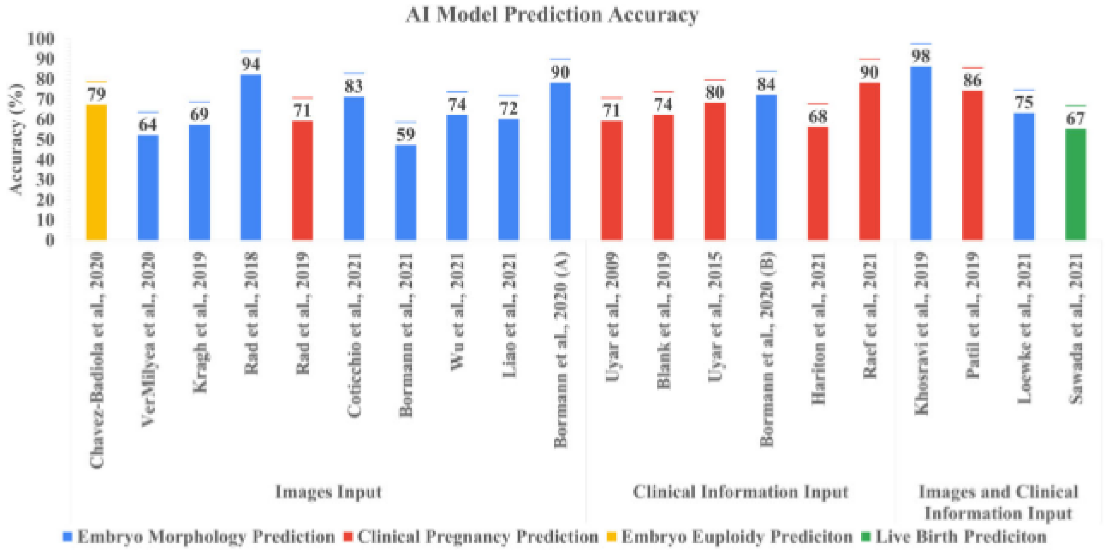


Figure 2. Accuracy of the AI model defined by the data input utilized for model training. Images input: studies including still embryo images and embryo images from timelapse videos. Clinical information input: studies including patient information features, demographics, and treatment information. Images and clinical information input: studies including the use of both embryo images and clinical information. The graph shows the accuracy output of the prediction defined by each study's input sample type, such as embryo morphology prediction. Embryo grade though images and their quality assessments; clinical pregnancy prediction: prediction of possible successful clinical pregnancy. Embryo aneuploidy prediction: embryo aneuploidy prediction through embryo images. Live birth prediction: prediction of possible successful healthy live birth delivery. AI, artificial intelligence.

Figure 5: accuracy of the AI model defined by the data input utilized for model training.

9.CONCLUSION

According to recent research on AI's use in embryo selection, AI models can predict reproductive outcomes and evaluate embryo morphology more accurately than embryologists. It is crucial to emphasise that the current accuracies claimed by studies exhibit varying performances because to the variety of techniques utilised, sample size, and datasets used for AI validation and training. Determining the minimal features of AI models needed for clinical application, as well as the accuracy of datasets and AI benchmark performance, are crucial for validation in the IVF industry. The top prediction models for clinical outcomes integrated images and clinical data, suggesting that including clinical data could enhance the performance of algorithms that just employ time-lapse videos or photographs. The most clinically meaningful outcome of ART is a live birth, hence future research should concentrate on predicting live births.

A completely automated deep learning model, called iDAScore v1.0, was created using 115,832 time-lapse sequences of embryo development. For KID embryos, the selection model's AUC is 0.67.

This completely automated model was shown to outperform the most advanced morphokineticKIDScore D5 v3 model. The high performance was attained without the embryologist's evaluation or annoyance.

REFERENCES

1. .Pribenszky C, Ma ´tya ´s S, Kova ´cs P, Losonczi E, Za ´dori J, Vajta G. Pregnancy achieved by transfer of a single blastocyst selected by time-lapse monitoring. Reproductive BioMedicine Online. 2010; 21 (4):533–6. https://doi.org/10.1016/j.rbmo.2010.04.015 PMID: 20638906

2. Ciray HN, Campbell A, Agerholm IE, Aguilar J, Chamayou S, Esbert M, et al. Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. Human Reproduction. 2014/10/24. 2014 Dec [cited 2020 Jan 13]; 29(12):2650–60. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25344070

3. Herrero J, Meseguer M. Selection of high potential embryos using time-lapse imaging: the era of mor- phokinetics. Fertility and sterility. 2013/02/06. 2013 Mar 15 [cited 2020 Jan 13]; 99(4):1030–4. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23395415.

4. Petersen BM, Boel M, Montag M, Gardner DK. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3. Human Reproduction. 2016/09/08. 2016 Oct 1 [cited 2020 Jan 13]; 31(10):2231–44. Available from: https:// www.ncbi.nlm.nih.gov/pubmed/27609980

5. Meseguer M, Herrero J, Tejera A, Hilligsøe KM, Ramsing NB, Remohı ´J, et al. The use of morphoki- netics as a predictor of embryo implantation. Human Reproduction. 2011/08/09. 2011 Oct [cited 2020 Jan 13]; 26(10):2658–71. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2182811

6. Milewski R, Ajduk A. Time-lapse imaging of cleavage divisions in embryo quality assessment. Repro- duction. 2017/04/13. 2017 Aug [cited 2020 Jan 13]; 154(2):R37–53. Available from: https://www.ncbi. nlm.nih.gov/pubmed/28408705

7. Ramsing NB, Berntsen J, Callesen H. Automated detection of cell division and movement in time-lapse images of developing bovine embryos can improve selection of viable embryos. Fertility and Sterility. 2007; 88:S38

8. Conaghan J, Chen AA, Willman SP, Ivani K, Chenette PE, Boostanfar R, et al. Improving embryo selec- tion using a computer-automated time-lapse image analysis test plus day 3 morphology: results from a prospective multicenter trial. Fertility and sterility. 2013/05/28. 2013 Aug [cited 2020 Jan 13];100(2):412–9.Availablefrom: https://www.ncbi.nlm.nih.gov/pubmed/23721712

9. Motato Y, de los Santos MJ, Escriba MJ, Ruiz BA, Remohı ´J, Meseguer M. Morphokinetic analysis and embryonic prediction for blastocyst formation through an integrated time-lapse system. Fertility and ste- rility. 2016 Feb [cited 2020 Jan 13]; 105(2):376–84.e9. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/26598211

10. Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Hickman CFL. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. Reproductive BioMedicine Online. 2013/02/19. 2013 May [cited 2020 Jan 13]; 26(5):477–85. Available from: https://www.ncbi.nlm.nih.gov/ pubmed/2351803

11. Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Thornton S. Retrospective analysis of outcomes after IVF using an aneuploidy risk model derived from time-lapse imaging without PGS. Reproductive biomedicine online. 2013/05/09. 2013 Aug [cited 2020 Jan 13]; 27(2):140–6. Available from: https:// www.ncbi.nlm.nih.gov/pubmed/23683847

12. Minasi MG, Colasante A, Riccio T, Ruberti A, Casciani V, Scarselli F, et al. Correlation between aneu- ploidy, standard morphology evaluation and morphokinetic development in 1730 biopsied blastocysts: a consecutive case series study. 2016/09/02. 2016 Oct [cited 2020 Jan 13]; 31(10):2245–54. Available from: https://academic.oup.com/humrep/article-lookup/doi/10.1093/humrep/dew183

13. Fishel S, Campbell A, Foad F, Davies L, Best L, Davis N, et al. Evolution of embryo selection for IVF from subjective morphology assessment to objective time-lapse algorithms improves chance of live birth. Reproductive biomedicine online. 2019 Oct 17;S1472-6483(19)30756-4. Available from: https:// www.ncbi.nlm.nih.gov/pubmed/318313

14. Rienzi L, Cimadomo D, Delgado A, Minasi MG, Fabozzi G, del Gallego R, et al. Time of morulation and trophectoderm quality are predictors of a live birth after euploid blastocyst transfer: a multicenter study. Fertility and sterility. 2019 Dec; 112(6):1080–93. Available from: https://www.ncbi.nlm.nih.gov/pubmed/ 31843084

15. Reignier A, Girard J-M, Lammers J, Chtourou S, Lefebvre T, Barriere P, et al. Performance of Day 5 KIDScore™morphokinetic prediction models of implantation and live birth after single blastocyst trans- fer. Journal of Assisted Reproduction and Genetics. 2019/08/23. 2019; 36(11):2279–85. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31444634

16. Pribenszky C, Nilselid A-MM, Montag M. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. Repro- ductiveBioMedicine Online. 2017/07/10. 2017 Nov [cited 2020 Jan 13]; 35(5):511–20. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1472648317303073

17. Magdi Y, Samy A, Abbas AM, Ibrahim MA, Edris Y, El-Gohary A, et al. Effect of embryo selection based morphokinetics on IVF/ICSI outcomes: evidence from a systematic review and meta-analysis of ran- domized controlled trials. Archives of gynecology and obstetrics. 2019/10/30. 2019 Dec; 300(6):1479– 90. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31667608

18. Barrie A, Homburg R, McDowell G, Brown J, Kingsland C, Troup S. Examining the efficacy of six pub- lished time-lapse imaging embryo selection algorithms to predict implantation to demonstrate the need for the development of specific, in-house morphokinetic selection algorithms. Fertility and Sterility. 2017/01/06. 2017 Mar 1 [cited 2020 Jan 13]; 107(3):613–21. Available from: https://linkinghub.elsevier. com/retrieve/pii/S0015028216630145.

19. Alikani M, Go KJ, McCaffrey C, McCulloh DH. Comprehensive evaluation of contemporary assisted reproduction technology laboratory operations to determine staffing levels that promote patient safety and quality care. Fertility and Sterility. 2014; 102(5):1350–6. Available from: http://dx.doi.org/10.1016/j. fertnstert.2014.07.1246

20. Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time- lapse annotations. Human Reproduction. 2013/09/26. 2013 Dec 1 [cited 2020 Jan 13]; 28(12):3215–21. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2407099

21. Adolfsson E, Andershed AN. Morphology vs morphokinetics: A retrospective comparison of interob- server and intra-observer agreement between embryologists on blastocysts with known implantation outcome. JornalBrasileiro de ReproducaoAssistida. 2018 Sep 1 [cited 2020 Jan 13]; 22(3):228–37. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29912521.

22. 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. Human Reproduction. 2017; 32(2):307–14. https://doi.org/10.1093/humrep/dew330 PMID: 28031323

23. Bormann CL, Ph D, Thirumalaraju P, Tech B, Kanakasabapathy K, Tech M. Consistency and objectivity of automated embryo assessments using deep neural networks. Fertility and sterility. 2020; 113 (4):781–7. https://doi.org/10.1016/j.fertnstert.2019.12.004 PMID: 32228880

24. Giusti A, Corani G, Gambardella L, Magli C, Gianaroli L. Blastomere segmentation and 3D morphology measurements of early embryos from hoffman modulation contrast image stacks. In: 2010 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2010—Proceedings. 2010. p. 1261–4.

25. Wang Y, Moussavi F, Lorenzen P. Automated embryo stage classification in Time-Lapse Micros- copy Video of Early Human Embryo Development. Medical image computing and computer-assisted intervention: MICCAI. International Conference on Medical Image Computing and Computer-Assis- ted Intervention. 2013; 16(Pt 2):460–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/ 24579173