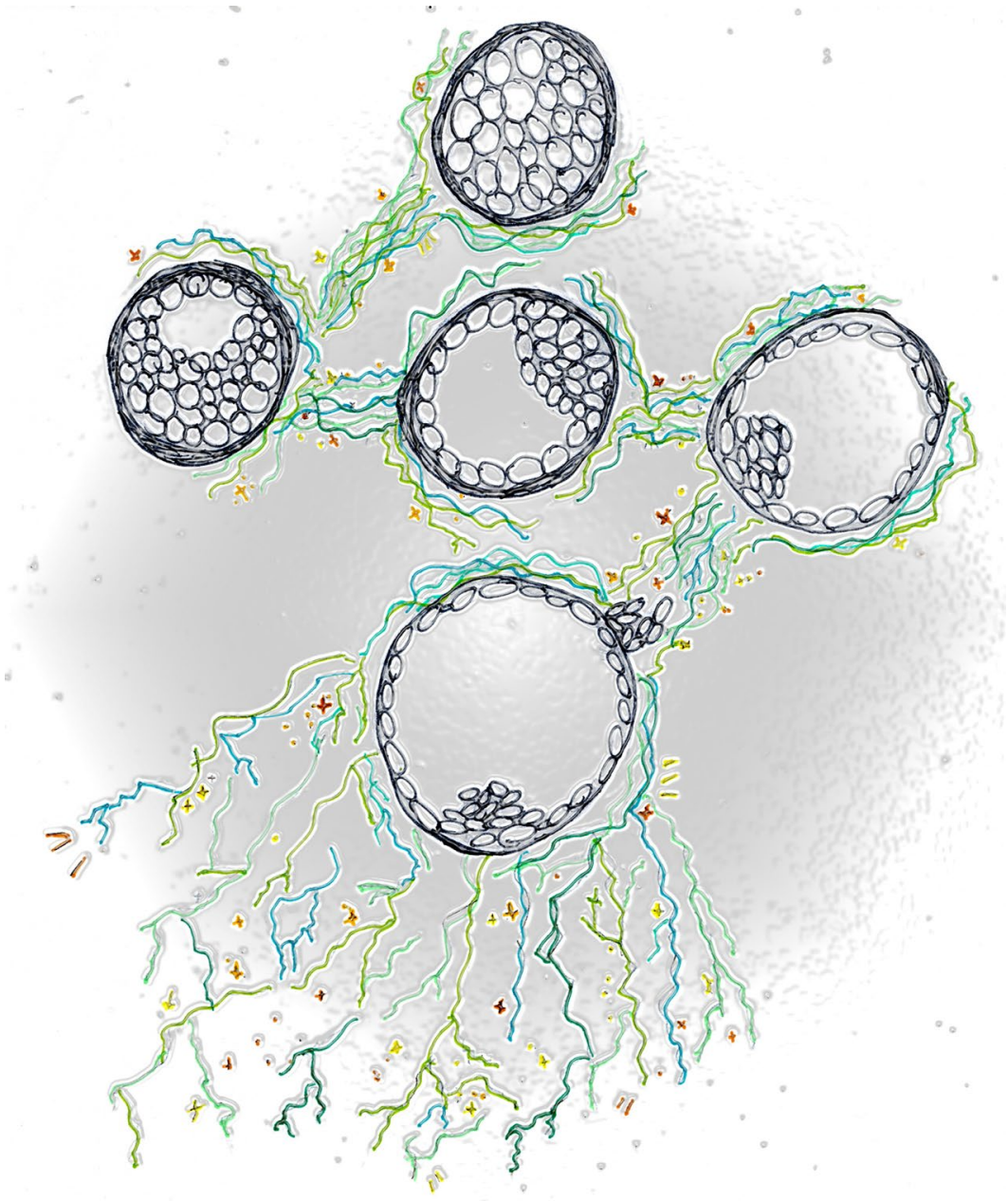
Predicting Blastocyst Viability:

A Machine Learning Approach Using Automated Blastocyst Expansion Measurements, Clinical Variables and Images



Koen Kwakkenbos

Supervisors:
dr. J.F. Veenland, dr. E.B. Baart, dr. M. Arif, drs. E.A. Chavli

Cover art: Julia Kwakkenbos

PREDICTING BLASTOCYST VIABILITY

-A Machine Learning Approach Using Automated Blastocyst Expansion Measurements, Clinical Variables and Images-

Koen Kwakkenbos

Student number: 4858956

09 Jul 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University; Delft University of Technology; Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)

Dept. of Biomechanical Engineering, TUDELFT

November 2023 – July 2024

Supervisor(s):

Dr. Jifke Veenland

Dr. Esther Baart

Dr. Muhammad Arif

Drs. Effrosyni Chavli

Thesis committee members:

Dr. Jifke Veenland, TU Delft (chair)

Dr. Esther Baart, Erasmus MC

Dr. Muhammad Arif, Erasmus MC

Dr. Melek Rousian, Erasmus MC

Dr. Carolien van Deurzen, Erasmus MC

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Preface

The completion of this thesis marks the end of my six years of studying. Throughout the bachelor Clinical Technology and the master Technical Medicine, I became increasingly interested in medical imaging and the techniques surrounding it. The courses in the bachelor and master laid a strong foundation in both medical and technical knowledge, revealing the vast potential of imaging technologies across various scales of biology. Surprisingly, when I approached Jifke for a master's thesis project, she introduced me to a new world, that of IVF. However, regardless of the difference in the medical specialism, this period made it clear that the same advanced imaging and artificial intelligence techniques used for complex anatomical structures could be applied to microscopic entities like blastocysts.

I would like to express my gratitude to Esther and Jifke, who supervised me throughout my final year. Despite your busy calendars, you always had time for our progress meetings and for provided multiple rounds of valuable feedback on the products of my thesis year. The combination of Jifke's technical view and Esther's clinical view, made this thesis a truly enjoyable experience and an exemplary technical medicine project.

Furthermore, I want to thank Arif for his technical insights and critical thinking on the thesis work. Moreover, I am particularly grateful to Eefje for her input in the project and all her efforts in swiftly completing our database.

Next, I extend my thanks to Jeroen and Kaylee, for allowing me to attend the embryo scoring mornings and for their insightful suggestions and discussions about the iDAScore and blastocyst expansion.

I am grateful to all Erasmus MC IVF technicians for their warm reception to the IVF laboratory and their willingness to show me their work and their assistance with side-projects to enhance data collection for this thesis. I also wish to acknowledge the medical staff of the Voortplantingscentrum for finding opportunities for me to join outpatient clinics and fertility procedures.

I would like to thank Melek Rousian and Carolien van Deurzen for their willingness to join this thesis committee.

Finally, my heartfelt thanks go to my family (Anja, Paul, Julia and Anne) and friends for their support throughout my thesis work and entire period of study, as well as for providing welcome distractions. I also would like to express my appreciation to Luuk for his support during this thesis and our refreshing coffee breaks.

Koen Kwakkenbos
Nieuwkoop
July 2024

Only the part marked as TM30004 – TM Master thesis should be graded.
TM30003 – TM Literature Study has been assessed separately.

TM30004 – TM Master Thesis

Predicting Blastocyst Viability: A Machine Learning Approach Using Automated Blastocyst Expansion Measurements, Clinical Variables and Images

1 ABSTRACT

Objective: To develop and evaluate a machine learning approach for predicting blastocyst viability using automated expansion measurements, clinical variables and image features.

Methods: A convolutional neural network was developed to automatically segment and measure blastocyst cross-sectional area from time-lapse images. We generated expansion curves and extracted features for 315 blastocysts. Various machine learning models were trained to predict biochemical and ongoing pregnancy using expansion, clinical and image-derived features. Model performance was evaluated using cross-validation and an unseen test set.

Results: The segmentation model achieved a Jaccard index of 97.6% on the validation set. Support vector machines using clinical and expansion features achieved the highest performance, with AUCs of 0.71 and 0.70 for predicting biochemical and ongoing pregnancy, respectively, on the test set. Blastocysts resulting in pregnancy expanded significantly faster and reached larger final cross-sectional areas compared to those that did not implant. Key predictive features included expansion rate and maternal age.

Conclusions: Automated quantification of blastocyst expansion dynamics combined with clinical variables enables prediction of implantation potential. Incorporating objective expansion metrics into embryo selection may enhance IVF success rates beyond traditional morphological grading systems.

Abbreviations: ***BCE*** *– Binary crossentropy;* ***BMI*** *– body mass index;* ***CNN*** *– Convolutional neural network;* ***CSA*** *– Cross-sectional area;* ***HFEA*** *– Human Fertilisation and embryology Authority;* ***ICM*** *– Inner cell mass;* ***ICSI*** *– Intracytoplasmic sperm injection;* ***IVF*** *– In-vitro fertilisation;* ***KNN*** *– K-Nearest Neighbour;* ***LR*** *– Logistic regression;* ***RD-UNet*** *– Residual Dilated U-Net;* ***RF*** *– Random forest;* ***ROC-AUC*** *– Area under the Receiver Operating Characteristics curve;* ***SHAP*** *– Shapley Additive exPlanations (SHAP);* ***SVM*** *– Support vector machine;* ***TE*** *– Trophectoderm;* ***tSB*** *– Time of initiation of blastulation;* ***XGB*** *– XGBoost;* ***ZP*** *– Zona pellucida.*

2 INTRODUCTION

It is estimated that 17.5% of all couples worldwide may experience subfertility, the inability to achieve a pregnancy for a period of 12 months (1). On of the most commonly used treatments for subfertility is in vitro fertilisation (IVF) (2). An IVF treatment cycle starts with gonadotropin stimulation of the ovaries. Mature oocytes are then harvested and fertilised. This fertilisation can be performed either by mixing with spermatozoa (IVF) or using intracytoplasmic sperm injection (ICSI). The embryos are cultured in the IVF laboratory under optimized conditions and usually one is then transferred into the uterus. In 2021, a total of 13683 IVF cycles were initiated in the Netherlands, resulting in a total of 5955 ongoing pregnancies (including both fresh embryo transfers and transfers after cryopreservation) (3). This accounted for approximately 3% of all babies born that year (4).

During the course of an IVF treatment, one or multiple blastocysts will have developed on the fourth day after insemination. These early-stage embryos consist of a protective glycoprotein coat, the zona pellucida (ZP), an outer layer of epithelial cells, the trophectoderm (TE), which encloses the fluid-filled blastocoel cavity and an inner cell mass (ICM). The selection of viable blastocysts for embryo transfer is a critical step in IVF. Accurate selection of embryos with the highest implantation potential is essential to increase the success rates of IVF while minimizing the risks associated with multiple pregnancies. Selection is performed based on the Gardner and Schoolcraft embryo grading system (5). This system consists of three distinct subgrades: the blastocyst developmental stage, the number of cells and compaction of the ICM and the number of cells and quality of the TE. The selection of embryos that receive good grades is more likely to lead to a sustained pregnancy (6, 7), and can eliminate the need of transferring multiple embryos to the uterus (8), preventing the exposure of the mother to the complications of a possible multiple pregnancy.

One limitation of the current grading system is the relatively high inter-observer variability (9-11). This is particularly due to the qualitative nature of the grading system and the lack of quantitative features. The highest inter-observer agreement was found for the part of the grade that describes blastocyst development (11). This subgrade (scale 1-6) denotes the state of the cavity formation and blastocyst expansion (defined as a gain of volume and a resulting thinning of the ZP). This part of the grade is also most strongly related to implantation potential (12-18). Still, grading blastocyst development lacks the precision needed for ranking embryos within the same grade category. This variability can affect the reliability of selecting the best embryo for transfer, potentially increasing the time-to-pregnancy.

Blastocyst development can be quantified by the cross sectional area (CSA) as a marker for the degree of expansion. With the emergence of digital microscopy and time-lapse imaging of developing embryos, it became possible to perform digital measurements at multiple timepoints without having to transfer the embryo out of its controlled culture environment. Using these techniques, research using manual area measurements has shown that the blastocyst's CSA before transfer was significantly correlated to implantation and pregnancy viability (19). Using manually measured CSAs at different time-points on time-lapse microscopy imaging, the rate of expansion was also found to correlate significantly with pregnancy (20) and also with blastocyst euploidy (21). While CSA measurements may contribute to a more objective transfer decision, these measurements add workload for the embryologist and are therefore not feasible in clinical practice, especially when determining the rate of expansion. Therefore, studies have been performed using tools to automatically measure the blastocyst's CSA, at any given timepoint in a time-lapse sequence. These tools enable measurement of the blastocyst's final CSA, as well as expansion dynamics and rate (by measuring the CSA at different timepoints in the time-lapse footage). These studies also found a significant positive correlation between both the rate, as well as the degree of blastocyst expansion and ongoing pregnancy (22) and

blastocyst euploidy (23). These studies were generally limited to the average expansion rate and final CSA. However, a theoretical advantage of automated CSA measurements of the entire developmental period is the automatic extraction of more advanced expansion and developmental characteristics. Furthermore, the previous studies only established a significant correlation between expansion and blastocyst viability, and did not provide insight into the predictive value of parameters derived from these automated CSA measurements with regards to blastocyst implantation potential and pregnancy outcome.

Therefore, in this study, we aim to predict the implantation potential of blastocysts by utilizing automated extraction of expansion measurements from time-lapse images, image and clinical features. By developing and evaluating an artificial intelligence-based segmentation model, we can quantify blastocyst expansion and derive expansion curves that reflect the changes in blastocyst CSA during development. These expansion parameters, combined with clinical and blastocyst image features, were used to train machine learning algorithms to predict biochemical and ongoing pregnancy outcomes. Through this approach, we seek to identify key parameters that can enhance embryo selection in clinical practice.

3 METHODS

3.1 STUDY DESIGN

Figure 1 outlines the design of this study. Data were collected from the EmbryoScope database and clinical records of the Erasmus MC, University Medical Centre Rotterdam. The study involved two cohorts: the segmentation cohort and the expansion cohort. The segmentation cohort was used to develop the blastocyst expansion measurement tool. This tool uses the concept of image segmentation to analyse blastocyst development. Image segmentation involves partitioning digital images into multiple segments to simplify their representation and make them more meaningful for analysis. This technique was applied to time-lapse images to accurately and automatically measure the CSA of the blastocysts, to generate blastocyst expansion curves and extract relevant features. These features were combined with image and clinical features, and used to train machine learning models aimed at the prediction of pregnancy outcomes for the expansion cohort. Finally, we retrospectively analysed the time-lapse footages of the blastocysts in the expansion cohort with the commercially available iDAScore v2.0 (Vitrolife, Göteborg, Sweden) algorithm. This software package assigned a score on the scale 1-9.9 to each blastocyst based on the entire time-lapse footage. The iDAScore results were used as a benchmark to evaluate the performance of our developed algorithm.

3.2 DATA COLLECTION

3.2.1 Ethical considerations

The data used in this study were retrospective and pseudonymized. The study was performed in accordance with the ethical standards described in the 1964 Declaration of Helsinki. All time-lapse images included in this study were obtained during routine clinical practice.

3.2.2 Segmentation cohort

Fifty blastocysts were randomly selected from the EmbryoScope database from the period February 2021 – September 2023. The time-lapse images of each of the blastocysts were exported. Blastocysts were included if images were available from the time of initiation of blastulation (tSB) until at least blastocyst expansion and ZP thinning, to ensure training data from all developmental periods of the

blastocyst. The tSB is defined as the first frame that shows initiation of a blastocoel cavity formation, according to the consensus guidelines (24).

3.2.3 Expansion cohort

This retrospective cohort study included blastocysts cultured and transferred at the Erasmus MC, University Medical Centre Rotterdam, in the period September 2023 until Januari 2024. All day-5 single fresh embryo transfer (SET) blastocysts were included if they were cultured in a time-lapse incubator (EmbryoScope, Vitrolife, Göteborg, Sweden). Blastocysts were excluded if there were problems with image acquisition (such as partial obstruction) or if the embryo did not yet reach tSB at the time of transfer. Clinical variables related to the mother, paternal subfertility and to the IVF treatment were collected as well. The two collected outcomes were biochemical pregnancy and ongoing pregnancy. Biochemical pregnancy was defined by a positive urinary β-hCG test 10 days after blastocyst transfer. Ongoing pregnancy was confirmed by the presence of a foetal heartbeat on an ultrasound at 12 weeks of gestation.

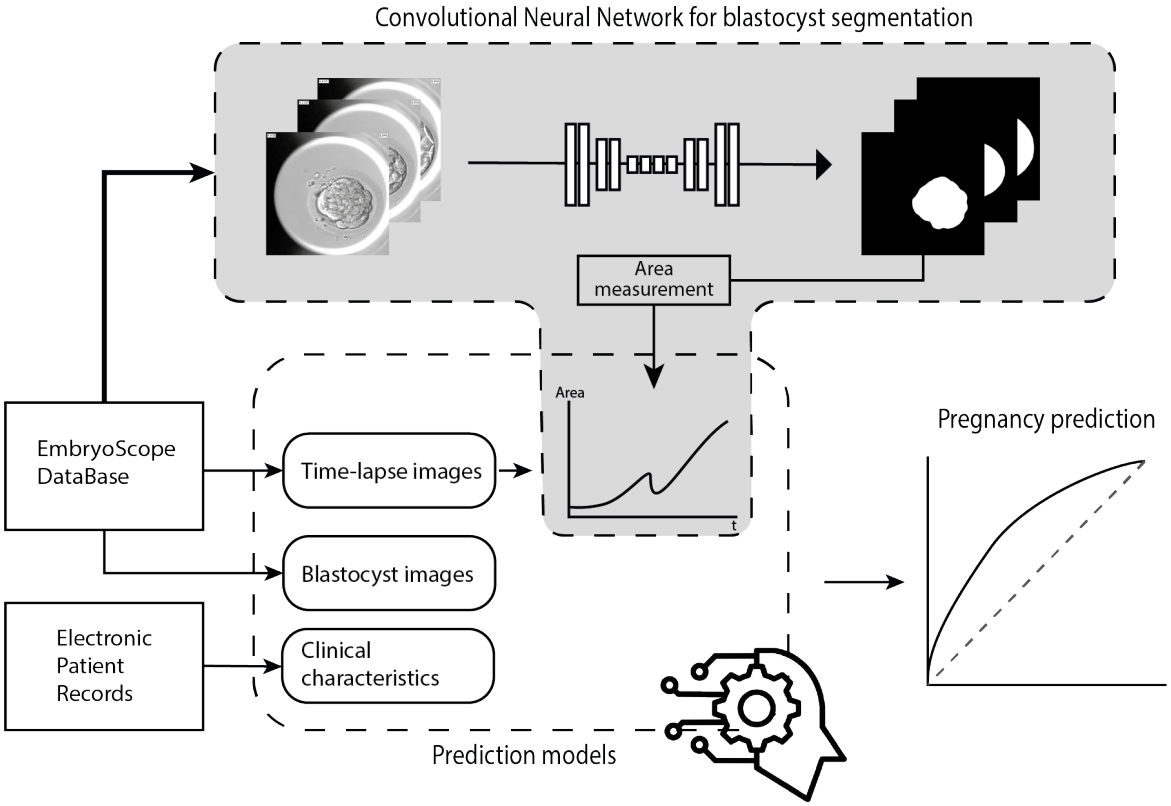


Figure 1: Schematic of the study design. Data from the EmbryoScope and electronic patient records were retrieved in the form of time-lapse images, static blastocyst images on day 5 and clinical characteristics of the included patients. The time-lapse images were used to generate expansion curves of the blastocysts. A convolutional neural network (CNN) for the automatic segmentation of the blastocyst cross-sectional area (CSA) was trained. Features from the expansion curves, combined with features extracted from images and clinical characteristics were used to train machine learning models aimed at prediction of biochemical and ongoing pregnancy.

3.3 IMAGE ACQUISITION AND EXPORT

The EmbryoScope records images in seven focal planes every 10 minutes. Images were extracted from tSB until the last available time-point. The tSB was manually identified by KK under supervision of an experienced embryologist (EB). Images were extracted in JPG format, with size 800x800 pixels. The middle focal plane was extracted (denoted by F0), as this often represented the largest area of cross-section through the embryo.

3.4 BLASTOCYST CSA QUANTIFICATION

From each of the 50 included time-lapse sequences, we extracted 10 evenly-spaced image frames in time, so that all developmental stages were represented. The blastocyst area (defined as the area of the embryo without the ZP) was manually annotated on each of the frames (in plane F0), using ImageJ software (25).

To automatically segment the CSA of the blastocyst, a convolutional neural network (CNN) was trained. A residual-dilated U-Net (RD-UNet) was used, which previously obtained good performance in the segmentation of blastocyst components and blastomeres (26-28). To increase the stability of the model, batch normalization layers were added between consecutive convolutional layers. The models were trained using different loss functions, batch sizes, and input normalization. The loss functions were minimized using the Adam optimizer (initial learning rate 0.001), and the models were trained for 100 epochs. Finally, to improve generalizability and to prevent overfitting on the training data, the models were trained with and without image augmentation. Augmentation steps included horizontal and vertical flipping, shifting the image by 10%, random rotation in the range [0, 270]°, random brightness and contrast altering, the addition of Gaussian noise and defocusing of the image. The models were implemented using Keras with a Tensorflow backend. The augmentation pipeline was implemented using the Albumentations Python package.

As the model outputs a continuous probability for each pixel, the masks were binarized by applying a threshold of 0.5. Finally, a postprocessing step selected the largest object in the mask and filled possible holes within this object.

To compare performances of different configurations, the models were trained using five-fold cross-validation. The dataset was split into five non-overlapping partitions at the blastocyst level, such that all time-lapse images of the same blastocyst were in the same fold. In each iteration, four partitions were used for training and the remaining one for validation, repeating this process five times. Model performance was measured using the average Jaccard Index across the validation splits, defined as: the size of the intersection divided by the size of the union of the predicted mask and the ground truth mask. The best-performing configuration, based on the highest average Jaccard Index, was retrained on the entire dataset. This model was thereafter tested on the publicly available dataset of Saeedi et al. (29) to assess the generalizability. This dataset contains 249 light microscopy images of human blastocysts, along with annotations of the blastocyst components (ICM, TE, ZP). We generated ground truth masks of the blastocyst's CSA by filling taking the inside of the ZP masks in de dataset of Saeedi et al.

3.5 BLASTOCYST EXPANSION CURVES

3.5.1 Expansion curve generation

The trained segmentation model was used in the continuation of this study as tool to measure the blastocyst's CSA. The model outputs the segmentation mask for each frame in a time-lapse sequence.

The number of pixels classified as embryo in the segmentation masks was multiplied by the area of one pixel (0.12 μm^2). The correctness of the segmentation was measured by the smoothness of the circumference of the mask. The mask at a time-point was rejected if the standard deviation of the curvature of the mask was higher than 0.08 (value empirically chosen based on inspection of the training data). The rejected mask was then not used for calculation of the CSA. Extraction of the CSA at each time-frame allowed for not only extraction of blastocyst expansion values and rates, but also for visualization of the full developmental process from the tSB until transfer. This included any potential blastocyst collapses, which are spontaneous contractions of the blastocoel cavity. This collapse may be caused by the efflux of blastocoel fluid as a result of loose cell bindings in the TE (30).

To standardize the analysis of blastocyst development, the time-lapse sequences of each embryo were realigned so that the tSB corresponds to t = 0 hours. This temporal realignment ensures that the expansion curves for all blastocysts begin at a consistent developmental stage, facilitating more accurate comparisons and population-level analyses. Due to the variability in the timing of blastulation among different embryos, the lengths of the expansion curves after realignment were not uniform. This resulted in a gradual reduction in the number of expansion curves available for averaging at each subsequent time point, potentially impacting the robustness of later-stage expansion data.

3.6 FEATURE EXTRACTION

Blastocyst expansion features were extracted from the time-lapse sequences. These included metrics related to the overall expansion, expansion rate, and characteristics of blastocyst collapses. A collapse was defined as a 10% reduction in the blastocyst's CSA. Table 1 provides an overview of the extracted parameters, with detailed definitions available in Table A.1 (Appendix A). Clinical parameters were extracted from electronic patient records as the second feature set. These included IVF treatment-related factors (such as insemination method, number of oocytes harvested, fertilised, and developed), patient characteristics (maternal age and Human Fertilisation and Embryology Authority (HFEA) age category), and the presence of male factor subfertility. Table 1 presents these clinical features alongside the expansion parameters. Finally, a pre-trained Xception CNN was used to extract image features. This architecture was chosen based on performance in blastocyst morphology classification in literature (31, 32).

Table 1: Overview of extracted expansion and clinical parameters. CSA: cross-sectional area, HFEA: Human Fertilisation and Embryology Authority; tSB: time of initiation of blastulation.

Expansion parameters	Clinical Parameters
Maximum CSA reached	Maternal age
Overall expansion	HFEA age category
Average expansion rate	Insemination method
Maximum expansion rate	Oocytes aspirated
Number of collapses	Number of fertilised zygotes
Average magnitude of collapses	Number of embryos used
Maximum magnitude of collapses	Embryo usage rate
Average time spend in collapse	Male factor
Maximum time spend in collapse	
tSB	
Developmental time	
Fitted exponential expansion rate	

3.7 Development of prediction models

We trained various machine learning algorithms, including Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbours (KNN), Random Forest (RF), and XGBoost (XGB) aimed at predicting biochemical pregnancy and ongoing pregnancy. The models were developed using the three distinct feature sets: clinical features, expansion pattern features, and image features and all possible combinations of the three. For the SVM, LR, and KNN models, continuous features were standardized using z-score normalization, so that the data was redistributed to have zero mean and unit variance. The scaling factors were calculated using the training data and applied to the validation and testing data. All categorical variables were one-hot encoded to ensure proper model training.

All models underwent training using five-fold cross-validation on the dataset collected from September to December 2023. The folds were stratified based on outcome, to ensure an equal distribution in each fold. Within each fold, the best model configuration was identified through hyperparameter optimization using the Hyperopt Python package (33), employing an inner five-fold cross-validation loop for this purpose. The search space for each model is available in Table B.1 (Appendix B). To address potential class imbalance issues, oversampling was applied using Synthetic Minority Over-sampling Technique (SMOTE) (34) on the training folds for predicting ongoing pregnancy. After determining the optimal model architecture and feature combination through cross-validation, we finalized the model by training it on the complete dataset from September to December 2023. This final model was then evaluated using the test data from January 2024. We calculated the area under the receiver operating characteristics curve (ROC-AUC) to evaluate the performance of the model. For comparative purposes, the ROC-AUC for the iDAScore v2.0 on the entire cohort (September 2023 – January 2024) was also calculated, similar to the work by Lassen et al. (35)

For both biochemical and ongoing pregnancy, we calculated Shapley Additive exPlanations (SHAP) values (36) for all samples in the five validation folds. These values were aggregated over all validation folds to gain insight in the most predictive features.

3.8 Statistical analysis

The baseline characteristics of the included blastocyst transfers were tested for the assumption of normality using the Shapiro-Wilk test. Statistical significance between the outcome groups was investigated using the Mann-Whitney U test. Categorical variables were analysed with the Chi-square test. Two-sided p-vales < 0.05 were considered statistically significant. Expansion curves of both outcomes (biochemical pregnancy and ongoing pregnancy) were compared using a permutation test of the difference between two groups of expansion curves (37).

4 Results

4.1 Patient characteristics

A total of 332 blastocyst transfers were available for analysis. From these, eight blastocysts were excluded due to partial obstruction of the images and nine as the blastocyst did not reach tSB. Finally, 255 blastocysts contributed to the training and validation set, while 60 were available in the test set (Figure 2). In total, 153 IVF, 136 ICSI treatments with ejaculated sperm and 26 TESE-ICSI treatments were included. Biochemical pregnancy was achieved in 163 transfers (51.7%) and 115 transfers (36.5%) resulted in an ongoing pregnancy. The baseline characteristics of the included blastocyst transfers are displayed in Table 1.

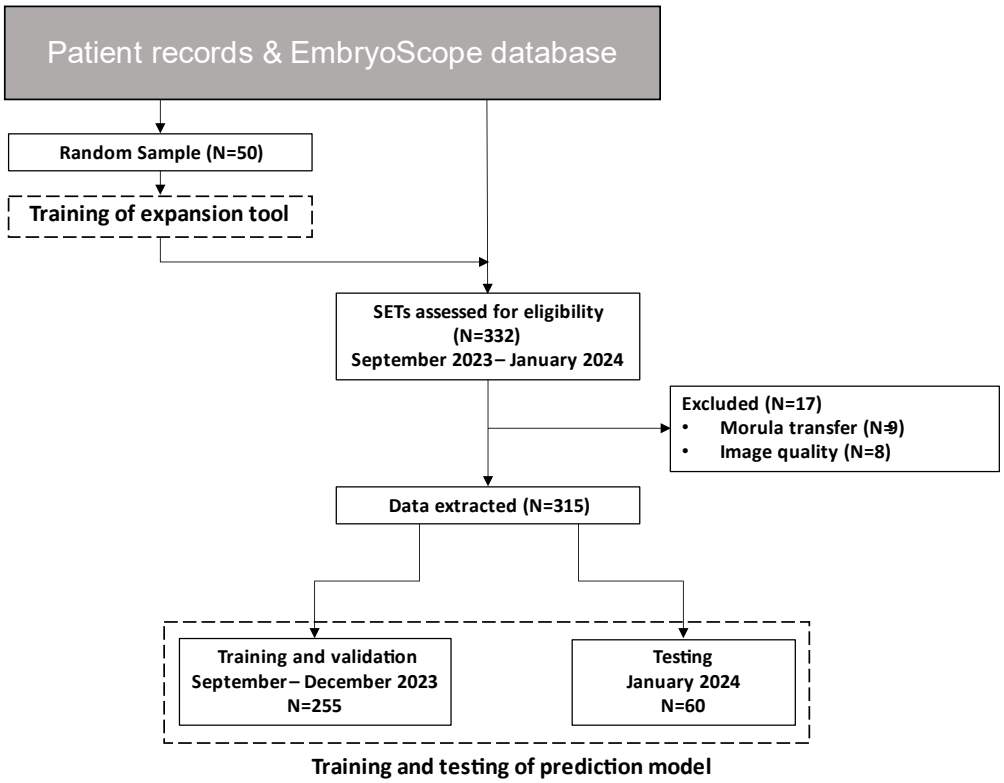


Figure 2: Flowchart showing blastocyst inclusion in this study. SET: single embryo transfer.

Table 2: Baseline characteristics per treatment outcome of the included blastocyst transfers. Data presented as median (interquartile range). HFEA: Human Fertilisation and Embryology Authority; IVF: in-vitro fertilization; ICSI: intracytoplasmic sperm injection.

	Biochemical pregnancy			Ongoing pregnancy		
	Implanted (N=132)	**No implantation (N=123)**	***p*-value**	**Pregnancy (N=94)**	**No pregnancy (N=161)**	***p*-value**
Female age	34.0 (30.0-37.0)	36.0 (33.0-39.0)	<0.001	34.0 (30.0-36.0)	36.0 (33.0-38.8)	<0.001
HFEA age category			<0.001			<0.001
18-34	62.0%	40.8%		63.7%	45.0%	
35-37	20.2%	26.3%		23.0%	23.3%	
38-39	12.9%	9.2%		9.7%	11.9%	
40-42	4.9%	23.0%		3.5%	19.3%	
43-44	0.0%	0.7%		0.0%	0.5%	
Oocytes aspirated	8.0 (6.0-12.0)	8.5 (6.0-12.0)	0.583	9.0 (6.0-12.0)	8.0 (6.0-12.0)	0.734
Male factor only	45.4%	25.0%	<0.001	48.7%	28.2%	<0.001
Embryo usage rate	0.60 (0.50-0.79)	0.51 (0.39-0.75)	0.013	0.67 (0.50-0.80)	0.56 (0.40-0.75)	0.005
Treatment			0.035			0.038
IVF	41.2%	55.9%		38.9%	54.0%	
ICSI (ejaculated sperm)	49.7%	36.2%		51.3%	38.6%	
TESE-ICSI	8.6%	7.9%		9.7%	7.4%	
iDAScore®	6.9 (5.0-7.9)	5.2 (2.8-7.6)	<0.001	6.8 (5.0-7.9)	6.0 (3.2-7.7)	0.002

4.2 BLASTOCYST QUANTIFICATION ALGORITHM

The explored configurations and their Jaccard index are provided in Table C.1 (Appendix C). The best performing segmentation model was the RD-UNet optimized with the Dice score loss function, achieving a Jaccard Index of 97.6% ± 0.58 (mean ± standard deviation). The performance on the external dataset was 98.0%.

4.3 BLASTOCYST EXPANSION CURVES AND FEATURES

The average expansion curves for the pregnancy outcomes are displayed in Figure 3, the derived parameters are shown in Table 3. Blastocysts that led to either a biochemical pregnancy or ongoing pregnancy reached a significantly larger CSA compared to those that did not (10481 µm^2 versus 6686 µm^2 and 10481 µm^2 versus 7256 µm^2, respectively). The average and maximum expansion rates were also significantly higher for blastocysts that led to a biochemical or ongoing pregnancy compared with those that did not. The difference in expansion rates is also visualized in the average expansion curves, particularly noticeable beyond 10 hours post-tSB, where the divergence between the two groups becomes more pronounced (Figure 3). The average expansion curves for both biochemical pregnancy and ongoing pregnancy were significantly different (*P*=0.035 and *P*=0.025, respectively). Noticeably, no significant differences were found between the number and magnitude of blastocyst collapses between the outcome groups (Table 3). Finally, the tSB was not significantly different between blastocysts that led to a biochemical or ongoing pregnancy and those that failed to do so.

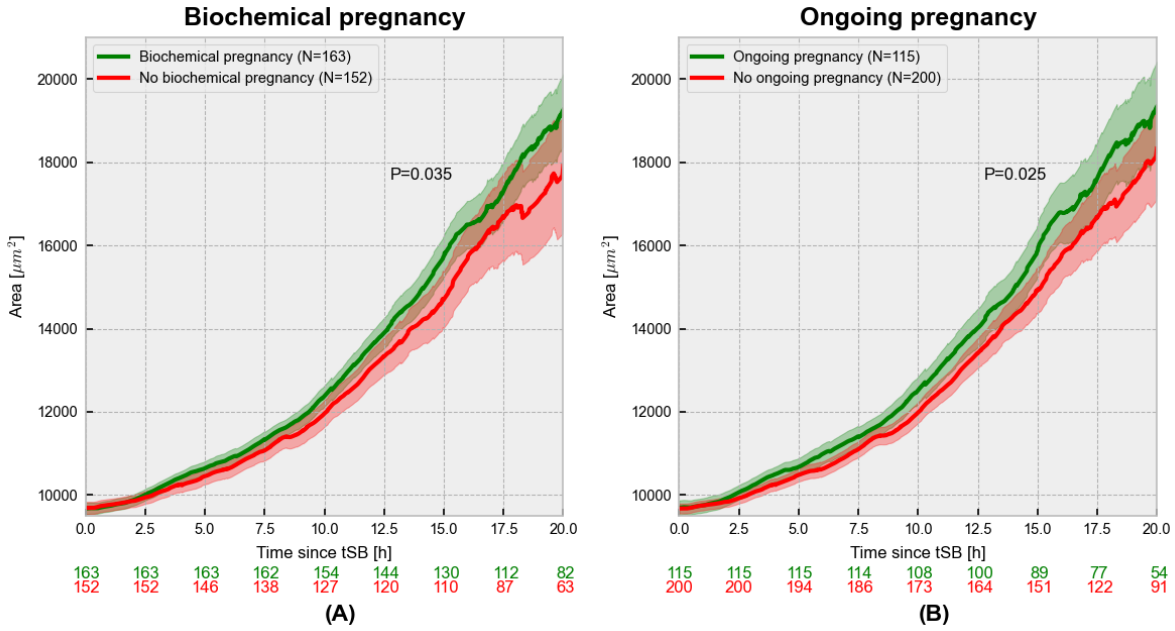


Figure 3: Average expansion curves for the included cohort. The curves show the average cross-sectional area (CSA) along with the 95% confidence intervals. Figure 3A displays the biochemical pregnancy outcome, while 3B shows the ongoing pregnancy. The starting point of the curves denotes the time of initiation of blastulation (tSB), as this absolute time-point differs for each blastocyst, blastocysts drop out of the analysis as time progresses. The number of blastocysts still in the analysis is shown at every 2.5 hours below the graphs. The colours correspond to those in the legend.

Table 3: Selection of measured parameters and their association with treatment outcome of the included cycles. Data are presented as median (interquartile range). P-values < 0.05 were considered significant (in bold). CSA: cross-sectional area; hpi: hours post insemination; tSB: time of initiation of blastulation.

	Biochemical pregnancy			Ongoing pregnancy		
	Positive (N=132)	Negative (N=123)	*p*-value	Pregnancy (N=94)	No pregnancy (N=161)	*p*-value
tSB [hpi]	95.50 (91.45-99.35)	96.65 (91.57-101.45)	0.086	96.10 (92.10-99.60)	96.00 (91.40-100.70)	0.633
Overall expansion [μm^2]	10480.64 (5338.87-13944.42)	6686.13 (3028.95-11414.78)	**<0.001**	10480.64 (5716.65-14041.70)	7256.15 (3754.66-12840.50)	**0.005**
Maximum cross-sectional area (CSA) [μm^2]	13314.06 (11763.07-14856.45	12219.78 (11007.53-13881.59)	**<0.001**	13372.6 (11929.9-14858.4)	12565.2 (11328.4-14151.7)	**0.019**
Average expansion rate [μm^2/h]	533.12 (362.44-653.56)	383.56 (233.75-566.60)	**<0.001**	535.92 (376.89-670.85)	404.76 (251.88-590.11)	**0.002**
Maximum expansion rate [μm^2/h]	2273.79 (1716.68-3040.77)	1953.76 (1424.39-2781.89))	**0.005**	2307.86 (1863.60-3097.22)	1971.61 (1480.13-2838.62)	**0.005**
Number of collapses	1.0 (0.0-1.0)	0.0 (0.0-1.0)	0.212	1.0 (0.0-1.0)	0.0 (0.0-1.0)	0.309
Average magnitude of collapses	0.26 (0.00-0.55)	0.00 (0.00-0.48)	0.174	0.25 (0.00-0.50)	0.00 (0.00-0.53)	0.415

4.4 BIOCHEMICAL AND ONGOING PREGNANCY PREDICTION

4.4.1 Model performance

To gain insight into which parameters hold the most predictive value for pregnancy outcome, we used various machine learning algorithms in combination with the three feature sets to generate prediction models. The results of all configurations are displayed in Table D.1 and D.2 (Appendix D). Table 4 shows the best result obtained using each combination of feature sets, and which model obtained that result. The SVM using the clinical and expansion features achieved the highest ROC-AUC (0.70 ± 0.02 and 0.68 ± 0.05 respectively). The ROC curves of these models are displayed in Figure 4. Notably, an SVM using only clinical features obtained the second-highest AUC for both biochemical pregnancy and ongoing pregnancy (0.68 and 0.66 respectively), but with a larger standard deviation (0.05). Using only the extracted features from the day-5 blastocyst images, a random forest achieved an AUC of 0.67 (± 0.03).

After identification of the best model and combination of features (expansion and clinical) during the cross-validation, two SVMs were trained on all available data from September to December 2023. These models achieved test performances of 0.71 and 0.70 for biochemical and ongoing pregnancy, respectively, on the January 2024 set. In comparison, the iDAScore v2.0 achieved a ROC-AUC of 0.65 and 0.62 (biochemical and ongoing pregnancy, respectively) on the entirety of the cohort.

Table 4: For each feature set combination, the best model is shown for both the biochemical pregnancy and ongoing pregnancy outcome. The area under the receiver operating characteristics curve (ROC-AUC) is displayed as average (standard deviation) over the five validation folds. LR: logistic regression; SVM: support vector machine; RF: random forest.

	Biochemical pregnancy		Ongoing pregnancy	
Feature sets	**Model**	**ROC-AUC**	**Model**	**ROC-AUC**
Clinical	SVM	0.68 (0.05)	SVM	0.66 (0.08)
Expansion	SVM	0.66 (0.05)	RF	0.59 (0.03)
Images	RF	0.67 (0.03)	XGBoost	0.56 (0.04)
Clinical & expansion	SVM	**0.70 (0.02)**	SVM	**0.68 (0.05)**
Clinical & images	LR	0.66 (0.04)	LR	0.57 (0.02)
Images & expansion	LR/XGBoost	0.65 (0.04)	KNN	0.55 (0.04)
Clinical & expansion & images	LR	0.67 (0.04)	LR	0.56 (0.04)

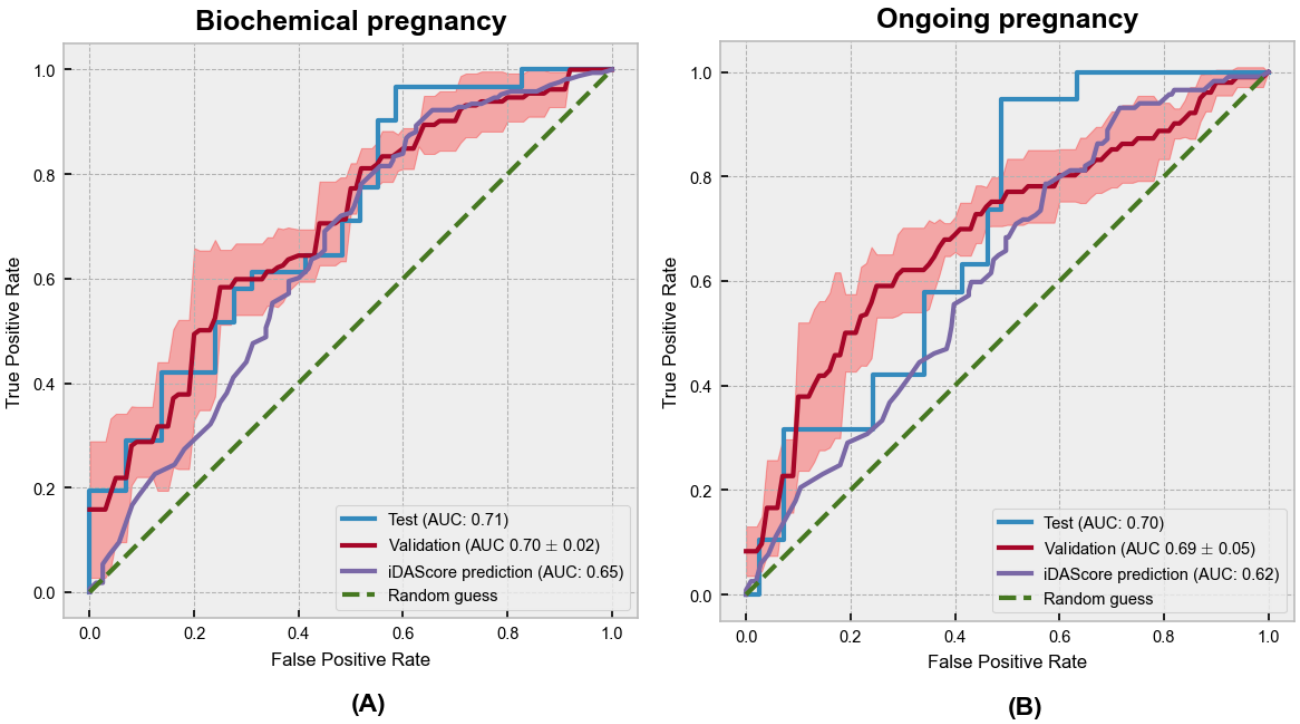


Figure 4: Receiver operating characteristics (ROC) curves for prediction of biochemical pregnancy (A) and ongoing pregnancy (B). The three curves in each plot show the results of the classifier on the test data, the cross-validation performance and the predictions of the iDAScore v2.0 on the full cohort. AUC: area under the curve.

4.4.2 Feature importance

The SHAP value plot depicted in Figure 5 illustrates the impact of various features on the prediction model's output for embryo implantation success. Each point on the plot represents a SHAP value for a specific feature, with the colour gradient indicating the feature's value (high values in red and low values in blue). Features are ordered by their overall importance in the model. The binary indicator of whether a blastocyst reached an expansion rate of more than 1000 µm²/hour was positively associated

with biochemical and ongoing pregnancy. Age was found to be the second most significant predictor of model output. Higher age values had a strong negative impact on implantation success.

The SHAP value analysis for ongoing pregnancy prediction revealed several key similarities and differences compared to the biochemical pregnancy prediction. In both models, age was an influential feature, consistently showing that as age increases this negatively influences pregnancy outcomes. The feature indicating a blastocoel expansion rate over 1000 µm²/hour also had a positive influence in both predictions, underscoring the importance of rapid expansion.

However, there were distinct differences in the relative importance of other features. For ongoing pregnancy prediction, the first HFEA age category (ages 18-34) showed a significant positive impact, indicating that younger age categories have a stronger association with ongoing pregnancies. The relative time to reach the maximum expansion rate, also displays a mixed but generally positive influence, suggesting its relevance in the ongoing pregnancy context. Insemination by means of ICSI was also found to be a significant predictor for ongoing pregnancy.

Interestingly, the average expansion rate and overall blastocyst expansion showed more importance in biochemical pregnancy prediction than in ongoing pregnancy. This suggests that while rapid and consistent early development is associated with implantation success, other factors may become more critical for maintaining pregnancy. Features such as number of oocytes aspirated and tSB (time of blastulation start) show varied impacts between the two predictions, indicating nuanced differences in the factors that influence biochemical versus ongoing pregnancy outcomes.

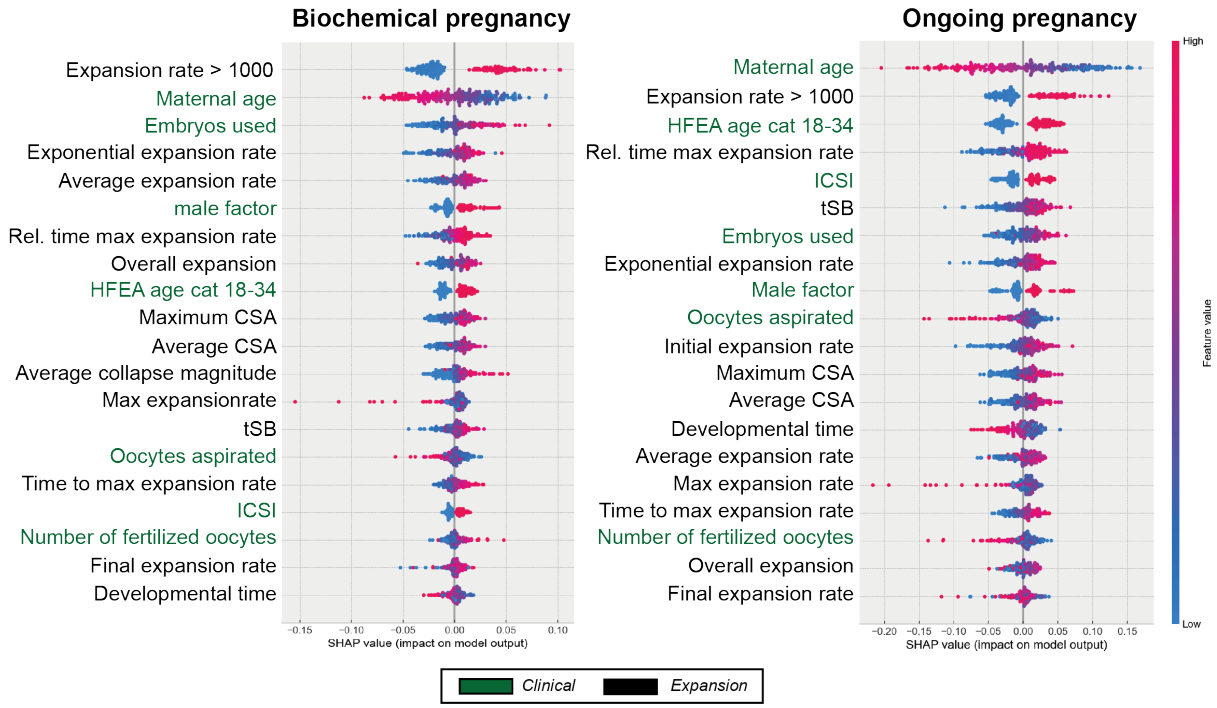


Figure 5: Shapley Additive exPlanations (SHAP) values. The clinical feature names are displayed in green, while expansion features are displayed in black. The position of each dot relative to the vertical axis (at 0.0) shows the impact on the model's prediction. CSA: cross-sectional area; tSB: time of initiation of blastulation; ICSI: intracytoplasmic sperm injection; HFEA: Human Fertilisation and Embryology Authority.

We also analysed feature importance specifically aimed at prediction of miscarriage (biochemical pregnancy without ongoing pregnancy). The SHAP values for miscarriage, shown in Figure E.1 (Appendix E), demonstrated similar patterns to those observed for ongoing pregnancy prediction. Given the similarity, we have included the figure in the appendix for further reference.

5 DISCUSSION

In this study, we developed and evaluated a CNN-based segmentation model to quantify blastocyst expansion from time-lapse images. We further explored the potential of various machine learning algorithms to predict biochemical and ongoing pregnancies based on blastocyst expansion, clinical and image-derived features, where the optimal model correlated with biochemical and ongoing pregnancy outcomes with AUCs of 0.71 and 0.70, respectively.

The results demonstrate that for quantifying blastocyst CSA, the RD-UNet model achieved high segmentation accuracy with a mean Jaccard Index of 97.6% on the validation sets and 98.0% on the external set of Harun et al. (26). This result is somewhat higher than the Jaccard index of 96.9% reported in Harun et al. who used a similar methodology (26). However, that study trained the model on a set of zona-ablated blastocysts, which have more irregular shapes. The size of the dataset in the study of Harun et al. is comparable to ours (617 versus 500 images), and both studies support the feasibility of training a CNN to automatically segment the embryo cross-sectional area. It is worth noting the relatively small variability in Jaccard Indices across different model configurations, as shown in Appendix C. The narrow range of performance (97.453% to 97.565%) suggests that the segmentation model is robust to various choices for loss functions and data preprocessing techniques for segmenting the CSA of a blastocyst.

The subsequent blastocyst expansion curves generated using the segmentation model enabled insights into blastocyst development. The curves for the blastocysts in both outcome groups differed significantly. The results in this work are in-line with previous findings acquired through manual blastocyst area annotation in our medical centre. Van Marion et al. found that embryos resulting in an ongoing pregnancy expanded significantly faster than those that did not (difference 100.9 μm^2/hour) (20). This value is comparable to the difference observed between the average expansion rates in our study (Table 3). Furthermore, our findings align with previous studies using automated blastocyst expansion measurements. For instance, Huang et al. (22) reported automatically generated expansion curves with embryos resulting in pregnancy demonstrating significantly greater average expansion rates, consistent with our observations.

The models predicting biochemical pregnancy generalized well, as indicated by the relatively low standard deviation around the AUC of 0.70 on the validation set. The AUC obtained on the unseen test data was similar to that of the cross-validation training. While other feature sets reached AUCs close to that of the best performing model, the combination of expansion curve features and clinical features seemed to have had a complementary effect reducing the variance of the models. Image features as stand-alone set achieve reasonable performance (AUC 0.67) for expansion, but performed worse for pregnancy (AUC 0.59). Features extracted from the images may indirectly be related to the expansion of the blastocyst, i.e. the shape of the cells might change in an expanded blastocyst. However, in a model not including clinical features, the predictive value for ongoing pregnancy was reduced. Including all three feature sets (expansion, images and clinical features) in a model did not achieve the highest performance. This might have been caused by the model receiving too many features to learn from. The iDAScore algorithm achieved a ROC-AUC of 0.65 and 0.62 for biochemical and ongoing pregnancy, respectively. Notably, our SVM model using clinical and expansion features demonstrated

improved performance, with ROC-AUCs of 0.70 and 0.68 for biochemical and ongoing pregnancy, respectively. This suggests that explicitly incorporating expansion dynamics and relevant clinical factors may offer additional predictive power beyond what is captured by standard time-lapse imaging analysis alone.

The findings from the SHAP value analysis highlighted the nuanced differences in factors influencing biochemical and ongoing pregnancy predictions. Age was a crucial determinant in both models, reflecting its well-documented impact on fertility outcomes. Notably, the greater significance of average expansion rate and overall blastocyst expansion in biochemical pregnancy prediction suggests that while these expansion metrics are predictive for implantation, they might be less predictive for sustained pregnancy. This shift highlights the complexity of pregnancy maintenance, where female characteristics and developmental factors became more prominent in the models. Factors involved in miscarriage risk are likely to have an important impact on ongoing pregnancy prediction. Indeed, the increased importance of female age, as evidenced by the high predictive value of both age and the first HFEA age category, confirms this. Moreover, the predictive significance of ICSI could be explained by the younger population associated with this fertilisation method.

These findings suggest that while blastocyst expansion metrics may be useful predictors for implantation, a broader set of factors must be considered for predicting ongoing pregnancy success. For instance, while research has shown that blastocyst euploidy is also correlated with expansion (21, 23), it is also known that some chromosomal compositions exhibit similar developmental patterns to euploid blastocysts. These blastocysts can however develop different from euploid blastocysts after implantation, leading to loss of the embryo (38). However, blastocysts that have a normal chromosomal constitution (as determined by cell biopsies) may still fail to implant. Implantation failure or failure to achieve ongoing pregnancy might be caused by lifestyle and nutrition or the uterine and blastocyst-endometrial dialogue (39), factors which are currently not included in our prediction models.

Still, embryo implantation is the first crucial step in achieving a sustained pregnancy. Therefore, the task of the embryologist is to select the most viable embryo. The practical application of these models emphasizes the need to focus on embryo-specific factors since clinical factors are constant within the same patient. When presented with multiple blastocysts to select from, the blastocyst with the highest average expansion rate and overall expansion should be prioritized. Incorporating these features into the blastocyst selection process may overcome some of the limitations of the currently used Gardner and Schoolcraft system. Blastocyst expansion and expansion rate provide objective and repeatable measurements, which also allow for a ranking of blastocysts with the same Gardner score.

Bias may have arisen from the fact that the study was retrospective and embryos included in this study were selected for transfer by the embryologists. The embryos included in this study were those that received the highest grades, based on conventional morphological criteria. Therefore, it is unclear what role TE and ICM quality plays in predicting blastocyst implantation and ongoing pregnancy. The selection process inherently filters out embryos with lower grades, which may exhibit similar expansion rates, but have different implantation potential. Another limitation of the current study is the lack of inclusion of other clinical variables such as body mass index (BMI) and smoking status. These variables could potentially help predict miscarriage (2, 39). However, due to inclusion of patients from multiple external hospitals it was not feasible to complete our dataset with these additional variables.

Future research should focus on ranking blastocysts with respect to their expansion rate and cross-sectional area. These rankings should be compared in a retrospective fashion with the embryologists embryos selected for transfer and cryopreservation. Furthermore, the relation between parameters

extracted from the expansion curves and the currently used Gardner and Schoolcraft grading system should be established, to investigate whether an objective method of ranking embryos could complement this existing system, providing a more robust framework for embryo selection. Expanding the dataset with additional lifestyle factors (such as BMI and smoking status) and TE and ICM quality could enhance the predictive accuracy of the models. Also, this could help to gain insights in the relative importances of blastocyst morphology (TE and ICM quality) and the expansion parameters. Finally, the prediction models should be tested on an independent cohort from an external IVF laboratory to establish the generalizability.

In this study, we developed and validated a CNN-based segmentation model to automatically quantify blastocyst expansion from time-lapse images, achieving high segmentation performance. The generated expansion curves, combined with clinical and image-derived features, enabled effective prediction of biochemical and ongoing pregnancies, with AUC values of 0.71 and 0.70, respectively. Blastocysts that lead to either a biochemical or ongoing pregnancy expand significantly faster and reach a larger final CSA compared to blastocysts that did not. Our findings suggest that incorporating objective blastocyst expansion metrics, such as expansion rate and overall expansion, into the embryo selection process can enhance IVF success rates beyond traditional grading systems.

6 Code availability

The code to reproduce the experiments described in this work is available from: https://github.com/KoenKwakkenbos/BlastoAI

7 Acknowledgements

The authors wish to thank Eva de Lange (Delft University of Technology) for her work on developing part of the image segmentation pipeline and Patricia Duis von Damm (IVF laboratory Erasmus MC) for retrospectively collecting the iDAScores.

8 References

1. Cox CM, Thoma ME, Tchangalova N, Mburu G, Bornstein MJ, Johnson CL, et al. Infertility prevalence and the methods of estimation from 1990 to 2021: a systematic review and meta-analysis. Hum Reprod Open. 2022;2022(4):hoac051.
2. Carson SA, Kallen AN. Diagnosis and Management of Infertility: A Review. Jama. 2021;326(1):65-76.
3. Landelijke IVF-cijfers 2021 [Internet]. 2021. Available from: https://www.degynaecoloog.nl/wp-content/uploads/2024/03/IVFlandelijk2021.pdf.
4. Centraal Bureau voor de Statistiek. Geboorte; kerncijfers, 1950-2022. 2023.
5. Gardner DK. Towards reproductive certainty: infertility and genetics beyond 1999. (No Title). 1999:378.
6. Bouillon C, Celton N, Kassem S, Frapsauce C, Guerif F. Obstetric and perinatal outcomes of singletons after single blastocyst transfer: is there any difference according to blastocyst morphology? Reprod Biomed Online. 2017;35(2):197-207.
7. Oron G, Son WY, Buckett W, Tulandi T, Holzer H. The association between embryo quality and perinatal outcome of singletons born after single embryo transfers: a pilot study. Hum Reprod. 2014;29(7):1444-51.

8. Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. Fertil Steril. 2000;73(6):1155-8.
9. Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: a multicenter study. Hum Reprod. 2017;32(2):307-14.
10. Cimadomo D, Sosa Fernandez L, Soscia D, Fabozzi G, Benini F, Cesana A, et al. Inter-centre reliability in embryo grading across several IVF clinics is limited: implications for embryo selection. Reprod Biomed Online. 2022;44(1):39-48.
11. Adolfsson E, Andershed AN. Morphology vs morphokinetics: a retrospective comparison of inter-observer and intra-observer agreement between embryologists on blastocysts with known implantation outcome. JBRA Assist Reprod. 2018;22(3):228-37.
12. della Ragione T, Verheyen G, Papanikolaou EG, Van Landuyt L, Devroey P, Van Steirteghem A. Developmental stage on day-5 and fragmentation rate on day-3 can influence the implantation potential of top-quality blastocysts in IVF cycles with single embryo transfer. Reprod Biol Endocrinol. 2007;5:2.
13. Thompson SM, Onwubalili N, Brown K, Jindal SK, McGovern PG. Blastocyst expansion score and trophectoderm morphology strongly predict successful clinical pregnancy and live birth following elective single embryo blastocyst transfer (eSET): a national study. J Assist Reprod Genet. 2013;30(12):1577-81.
14. Kresowik JD, Sparks AE, Van Voorhis BJ. Clinical factors associated with live birth after single embryo transfer. Fertil Steril. 2012;98(5):1152-6.
15. Du Q-Y, Wang E-Y, Huang Y, Guo X-Y, Xiong Y-J, Yu Y-P, et al. Blastocoele expansion degree predicts live birth after single blastocyst transfer for fresh and vitrified/warmed single blastocyst transfer cycles. Fertility and Sterility. 2016;105(4):910-9.e1.
16. Subira J, Craig J, Turner K, Bevan A, Ohuma E, McVeigh E, et al. Grade of the inner cell mass, but not trophectoderm, predicts live birth in fresh blastocyst single transfers. Human Fertility. 2016;19(4):254-61.
17. Van den Abbeel E, Balaban B, Ziebe S, Lundin K, Cuesta MJ, Klein BM, et al. Association between blastocyst morphology and outcome of single-blastocyst transfer. Reprod Biomed Online. 2013;27(4):353-61.
18. Storr A, Bilir E, Cooke S, Garrett D, Venetis CA. Fine-tuning blastocyst selection based on morphology: a multicentre analysis of 2461 single blastocyst transfers. Reprod Biomed Online. 2019;39(4):588-98.
19. Sciorio R, Thong D, Thong KJ, Pickering SJ. Clinical pregnancy is significantly associated with the blastocyst width and area: a time-lapse study. J Assist Reprod Genet. 2021;38(4):847-55.
20. van Marion ES, Chavli EA, Laven JSE, Steegers-Theunissen RPM, Koster MPH, Baart EB. Longitudinal surface measurements of human blastocysts show that the dynamics of blastocoel expansion are associated with fertilization method and ongoing pregnancy. Reprod Biol Endocrinol. 2022;20(1):53.
21. Park JK, Jeon Y, Bang S, Kim JW, Kwak IP, Lee WS. Time-lapse imaging of morula compaction for selecting high-quality blastocysts: a retrospective cohort study. Arch Gynecol Obstet. 2024;309(6):2897-906.
22. Huang TTF, Kosasa T, Walker B, Arnett C, Huang CTF, Yin C, et al. Deep learning neural network analysis of human blastocyst expansion from time-lapse image files. Reprod Biomed Online. 2021;42(6):1075-85.
23. Hori K, Hori K, Kosasa T, Walker B, Ohta A, Ahn HJ, et al. Comparison of euploid blastocyst expansion with subgroups of single chromosome, multiple chromosome, and segmental aneuploids using an AI platform from donor egg embryos. J Assist Reprod Genet. 2023;40(6):1407-16.
24. 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;29(12):2650-60.

25. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nature Methods. 2012;9(7):671-5.
26. Harun MY, Rahman MA, Mellinger J, Chang W, Huang TTF, Walker B, et al. Image Segmentation of Zona-Ablated Human Blastocysts. 2019 IEEE 13th International Conference on Nano/Molecular Medicine & Engineering (NANOMED). 2019:208-13.
27. Rad RM, Saeedi P, Au J, Havelock J. Blastomere Cell Counting and Centroid Localization in Microscopic Images of Human Embryo. 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)2018. p. 1-6.
28. Harun MY, Huang T, Ohta AT. Inner Cell Mass and Trophectoderm Segmentation in Human Blastocyst Images using Deep Neural Network. Ieee Int Conf Nano. 2019:214-9.
29. Saeedi P, Yee D, Au J, Havelock J. Automatic Identification of Human Blastocyst Components via Texture. IEEE Trans Biomed Eng. 2017;64(12):2968-78.
30. Marcos J, Pérez-Albalá S, Mifsud A, Molla M, Landeras J, Meseguer M. Collapse of blastocysts is strongly related to lower implantation success: a time-lapse study. Human Reproduction. 2015;30(11):2501-8.
31. Bormann CL, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, Kandula H, et al. Performance of a deep learning based neural network in the selection of human blastocysts for implantation. Elife. 2020;9.
32. Thirumalaraju P, Kanakasabapathy MK, Bormann CL, Gupta R, Pooniwala R, Kandula H, et al. Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality. Heliyon. 2021;7(2):e06298.
33. Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt: a Python library for model selection and hyperparameter optimization. Computational Science & Discovery. 2015;8(1).
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Int Res. 2002;16(1):321–57.
35. Theilgaard Lassen J, Fly Kragh M, Rimestad J, Nygård Johansen M, Berntsen J. Development and validation of deep learning based embryo selection across multiple days of transfer. Sci Rep. 2023;13(1):4235.
36. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.
37. Baldwin T, Sakthianandeswaren A, Curtis JM, Kumar B, Smyth GK, Foote SJ, et al. Wound healing response is a major contributor to the severity of cutaneous leishmaniasis in the ear model of infection. Parasite Immunol. 2007;29(10):501-13.
38. Shahbazi MN, Wang T, Tao X, Weatherbee BAT, Sun L, Zhan Y, et al. Developmental potential of aneuploid human embryos cultured beyond implantation. Nature Communications. 2020;11(1):3987.
39. Cimadomo D, Rienzi L, Conforti A, Forman E, Canosa S, Innocenti F, et al. Opening the black box: why do euploid blastocysts fail to implant? A systematic review and meta-analysis. Human Reproduction Update. 2023;29(5):570-633.

APPENDIX A: EXTRACTED FEATURES

Table A.1: Definitions and descriptions of extracted expansion parameters used in blastocyst analysis.

Expansion parameters	
Maximum CSA reached	Maximum cross-sectional area in the final 10 frames of the time-lapse footage, to compensate for any late collapse.
Overall expansion	Difference between maximum cross-sectional area in the final 10 frames of the time-lapse footage and the initial cross-sectional area.
Average expansion rate	Difference between maximum cross-sectional area in the final 10 frames of the time-lapse footage and the initial cross-sectional area divided by the developmental time.
Maximum expansion rate	Maximum value of the derivative of the expansion curve.
Number of collapses	Number of spontaneous blastocyst implosions > 10% of the current CSA.
Average magnitude of collapses	Average relative magnitude of the blastocyst implosions.
Maximum magnitude of collapses	Maximum magnitude of the observed implosions.
Average time spend in collapse	Average time (hours) to recover from all collapses.
Maximum time spend in collapse	Longest time taken to recover from collapse.
tSB	Time of initiation of blastulation.
Developmental time	Time from tSB until transfer.
Exponential expansion rate	Expansion rate of the best fitted exponential curve.

APPENDIX B: GRID SEARCH FOR MACHINE LEARNING MODELS

Table B.1: Hyperparameter search spaces for machine learning models used in pregnancy prediction. KNN: K-Nearest Neighbours; LR: Logistic regression; RF: Random forest; SVM: Support vector machine.

KNN	'n_neighbors': hp.uniformint('n_neighbors', 5, 30), 'weights': hp.choice('weights', ['uniform', 'distance']), 'algorithm': hp.choice('algorithm', ['auto', 'ball_tree']), 'leaf_size': hp.choice('leaf_size', np.arange(20, 50, 5)), 'p': hp.choice('p', [1, 2]),
LR	'C': hp.loguniform('C', -3, 0), 'penalty': hp.choice('penalty', ['l1', 'l2 'solver': hp.choice('solver', ['liblinear' 'tol': hp.loguniform('tol', -4, -2), 'max_iter': hp.choice('max_iter', [1000]),
RF	'n_estimators': hp.choice('n_estimators', np.arange(5, 150, dtype=int)), 'criterion': hp.choice('criterion', ['gini']), 'max_depth': hp.choice('max_depth', np.arange(1, 11, dtype=int)), 'min_samples_split': hp.choice('min_samples_split', np.arange(2, 11, dtype=int)), 'min_samples_leaf': hp.choice('min_samples_leaf', np.arange(1, 11, dtype=int)), 'min_weight_fraction_leaf': hp.uniform('min_weight_fraction_leaf', 0, 0.5), 'min_impurity_decrease': hp.uniform('min_impurity_decrease', 0, 0.5), 'class_weight': hp.choice('class_weight', [None, 'balanced']), 'n_jobs': hp.choice('n_jobs', [-1]), 'max_features': hp.choice('max_features', ['sqrt', 'log2', None])
SVM	'C': hp.loguniform('C', -2, 0), 'kernel': hp.choice('kernel', ['poly', 'rbf', 'sigmoid']), 'gamma': hp.choice('gamma', ['scale', 'auto']), 'coef0': hp.uniform('coef0', 0, 1), 'tol': hp.loguniform('tol', -4, -2), 'cache_size': hp.choice('cache_size', [2000]), 'shrinking': hp.choice('shrinking', [True, False]), 'break_ties': hp.choice('break_ties', [False, True]), 'class_weight': hp.choice('class_weight', [None, 'balanced']), 'probability': hp.choice('probability', [True]),
XGBoost	'eta': hp.loguniform('eta', -7, 0), 'max_depth': hp.choice('max_depth', np.arange(1, 11, dtype=int)), 'subsample': hp.uniform('subsample', 0.2, .5), 'colsample_bytree': hp.uniform('colsample_bytree', 0.2, .5), 'colsample_bylevel': hp.uniform('colsample_bylevel', 0.2, .5), 'min_child_weight': hp.loguniform('min_child_weight', -16, 2), 'alpha': hp.uniform('alpha', 0, 1), 'lambda': hp.uniform('lambda', 0, 1), 'gamma': hp.uniform('gamma', 0, 1), 'objective': 'binary:logistic',

APPENDIX C: SEGMENTATION PERFORMANCES

Table C.1: Performance comparison of various convolutional neural network configurations for blastocyst segmentation, measured by Jaccard index averaged over the five folds. BCE: binary crossentropy.

Loss function	Input Normalization	Augmentation	Jaccard index [%]
BCE	Min-max	No	97.524
BCE	Min-max	Yes	97.500
BCE	Batch normalization	No	97.565
BCE	Batch normalization	Yes	97.513
Focal BCE	Min-max	No	97.453
Focal BCE	Min-max	Yes	97.511
Focal BCE	Batch normalization	No	97.519
Focal BCE	Batch normalization	Yes	97.533
Dice	Min-max	No	97.539
Dice	Min-max	Yes	97.516
Dice	Batch normalization	No	97.561
Dice	**Batch normalization**	**Yes**	**97.565**
Weighted BCE dice	Min-max	No	97.533
Weighted BCE dice	Min-max	Yes	97.546
Weighted BCE dice	Batch normalization	No	97.534
Weighted BCE dice	Batch normalization	Yes	97.543

APPENDIX D: MACHINE LEARNING MODEL PERFORMANCES

Table D.1: Cross-validation performance (ROC-AUC) of machine learning models for predicting biochemical pregnancy using different feature combinations. Values displayed as mean (standard deviation) over the five folds. SVM: Support vector machine; KNN: K-Nearest Neighbour; RF: Random forest; LR: Logistic regression.

	Clinical	Curve	Image	Clinical + Curve	Clinical + Image	Curve + Image	Clinical + Curve + Image
KNN	0.67 (0.04)	0.62 (0.04)	0.63 (0.05)	0.66 (0.05)	0.64 (0.06)	0.62 (0.04)	0.63 (0.06)
LR	0.65 (0.04)	0.61 (0.06)	0.64 (0.05)	0.64 (0.05)	0.66 (0.04)	0.65 (0.04)	0.67 (0.04)
RF	0.65 (0.04)	0.61 (0.07))	0.67 (0.03)	0.62 (0.06)	0.65 (0.03)	0.64 (0.03)	0.64 (0.04)
SVM	0.68 (0.05)	0.66 (0.05)	0.63 (0.02)	**0.70 (0.02)**	0.63 (0.02)	0.63 (0.02)	0.64 (0.03)
XGBoost	0.65 (0.05)	0.62 (0.05)	0.65 (0.04)	0.67 (0.05)	0.62 (0.03)	0.65 (0.04)	0.63 (0.02)

Table D.2: Cross-validation performance (ROC-AUC) of machine learning models for predicting ongoing pregnancy using different feature combinations. Values displayed as mean (standard deviation) over the five folds. SVM: Support vector machine; KNN: K-Nearest Neighbour; RF: Random forest; LR: Logistic regression.

	Clinical	Curve	Image	Clinical + Curve	Clinical + Image	Curve + Image	Clinical + Curve + Image
KNN	0.62 (0.05)	0.56 (0.06)	0.53 (0.09)	0.66 (0.06)	0.53 (0.05)	0.55 (0.04)	0.53 (0.09)
LR	0.62 (0.08)	0.59 (0.04)	0.52 (0.03)	0.61 (0.08)	0.57 (0.02)	0.54 (0.05)	0.56 (0.04)
RF	0.63 (0.09)	0.59 (0.03)	0.55 (0.03)	0.60 (0.09)	0.49 (0.14)	0.52 (0.03)	0.47 (0.14)
SVM	0.66 (0.08)	0.59 (0.04)	0.51 (0.07)	**0.68 (0.05)**	0.50 (0.09)	0.52 (0.07)	0.51 (0.06)
XGBoost	0.63 (0.04)	0.58 (0.05)	0.56 (0.04)	0.63 (0.06)	0.53 (0.06)	0.55 (0.07)	0.54 (0.08)

APPENDIX E: FEATURE IMPORTANCE RELATED TO MISCARRIAGE

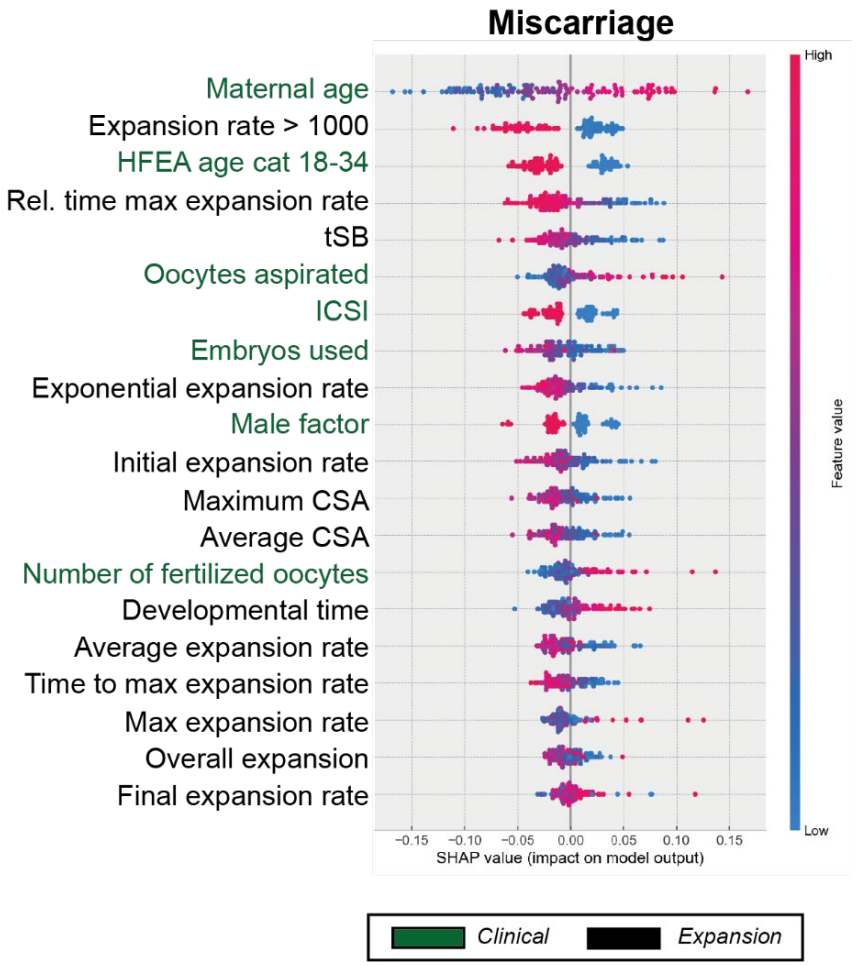


Figure E.1: Shapley Additive exPlanations (SHAP) values for miscarriage outcome. The clinical feature names are displayed in green, while expansion features are displayed in black. The position of each dot relative to the vertical axis (at 0.0) shows the impact on the model's prediction. CSA: cross-sectional area; tSB: time of initiation of blastulation; ICSI: intracytoplasmic sperm injection; HFEA: Human Fertilisation and Embryology Authority.

TM30003 – TM Literature Study

Advancements in Automated Image Quantification Techniques for the Human Blastocyst and Their Potential Clinical Relevance: A Systematic Review

1 ABSTRACT

The emergence of digital microscopy techniques potentially allows for automatic extraction of quantitative morphological features to aid the embryologist in their evaluation of blastocysts. To gain insight in the existing algorithms for automated blastocyst image quantification and their potential applications, this systematic review was performed. Twenty-two studies that described either the development of a blastocyst segmentation algorithm, or the application for clinical outcomes thereof, were included after a systematic search of Medline, Embase, Web of Science, Cochrane Central Register of Controlled Trials and Google Scholar. Accurate automated image quantification of blastocyst components was found to be feasible, especially using deep learning techniques. Only three studies explored the use of these algorithms in relation to clinical outcomes such as implantation and live birth. Blastocyst expansion, measured via automatic segmentation, was significantly correlated to live births and implantation. Future research should focus on the automated image quantification of blastocyst components during the full developmental process, as well as further exploring the predictive value of quantitative parameters for developmental potential.

Key message: Accurate automated image quantification of the blastocyst and its components is feasible. The focus in the field can now shift from the development of automated image quantification tools to the utilization of these tools for the prediction of clinical outcomes such as implantation and live birth.

Abbreviations: ***ANN*** *– artificial neural network;* ***BC*** *– blastocoel;* ***BG*** *– background;* ***CNN*** *– convolutional neural network;* ***ICM*** *– inner cell mass;* ***IVF*** *– in-vitro fertilisation;* ***PPV*** *– Positive predictive value;* ***SVM*** *– support vector machine;* ***TE*** *– trophectoderm;* ***ZP*** *– zona pellucida.*

2 INTRODUCTION

Subfertility impacts approximately 17.5% of the adult population worldwide (Cox et al. 2022). Among the various treatments available, in-vitro fertilization (IVF) stands out as one of the most commonly used. Usually, during the course of an IVF treatment, multiple blastocysts will develop and a clinical embryologist decides which blastocyst to transfer. The preference for a single embryo transfer makes selection of the embryo with the highest implantation potential crucial. This decision is currently based on the morphological quality of the blastocysts, as assessed using the Gardner and Schoolcraft embryo grading system (Gardner 1999). Blastocysts consist of four distinct components: the outer layer of cells, trophectoderm (TE), which surrounds the fluid filled cavity called the blastocoel (BC), the inner cell mass (ICM) or embryoblast, and the zona pellucida (ZP), which is the protective glycoprotein co at surrounding the embryo. The assigned grades are based on morphologic criteria relating to these components: highest grades for morphology are given if the blastocoel is fully expanded, the TE consists of many cells forming a cohesive epithelium and the ICM is made up of many cells compacted together (Gardner et al. 2000, Alpha Scientists in Reproductive and Embryology 2011, Thompson et al. 2013). Selection of embryos with higher grades is associated with higher chances of pregnancy and live birth (Oron et al. 2014, Bouillon et al. 2017) and can eliminate the need to transfer multiple blastocysts (Gardner et al. 2000). However, a limitation of the morphological grading is the high subjectivity (Storr et al. 2017). This assessment of blastocyst quality may be performed visually under a microscope, or using a digital image system. The latter option enables the use of image processing techniques to aid the embryologist in their decision, possibly contributing to a more objective clinical decision. This approach has taken a flight with the introduction of time-lapse incubators specifically designed for human IVF embryos into clinical practice, generating embryo images throughout in vitro development.

Identification of blastocyst components on digital images could enable extraction of quantitative parameters. Manual approaches have demonstrated relevance of quantitative parameters, such as morphological features (Lagalla et al. 2015) or blastocyst area and expansion (Lagalla et al. 2015, Eelbode et al. 2020, van Marion et al. 2022) in relation to clinical outcome. However, the manual extraction of those features is labour-intensive, and therefore not feasible for clinical practice. Image segmentation algorithms might be employed to automatically delineate blastocyst components. The identified segmentation masks may subsequently be used to extract such quantitative parameters.

Several image processing techniques aimed at automated image quantification or image segmentation exist. These range from traditional image processing methods, such as level-set algorithms and texture analysis that rely on mathematical methods to identify patterns and boundaries in images, to more contemporary deep learning techniques involving artificial- and convolutional neural networks (ANNs and CNNs, respectively). These ANNs and CNNs are algorithms that can automatically learn features directly from the images, given enough training examples. The aim of this systematic review is to comprehensively evaluate the state of technology regarding segmentation techniques applied to human blastocyst components, and their possible clinical application.

3 MATERIALS AND METHODS

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (Page et al. 2021), and its protocol has been registered in PROSPERO under registration ID CRD42024500712. This article adheres to ethical guidelines and did not require the use of informed consent.

3.1 SEARCH STRATEGY

The Medline, Embase, Web of Science, Cochrane Central Register of Controlled Trials and Google Scholar databases were searched from inception until 5-12-2023 (date last search). The general pattern of all queries was: Segmentation AND (Algorithm OR Deep Learning OR Machine Learning) AND (Blastocyst OR Embryo). The full strategies per database are reported in Appendix A.

3.2 INCLUSION CRITERIA AND STUDY SELECTION

After removal of duplicates, the titles and abstracts of articles were screened to assess eligibility for inclusion. Articles that described and reported performance measures of the segmentation of the blastocyst or a component (BC, ICM, TE or ZP) thereof on light microscopy images were included. Additionally, articles that applied a segmentation algorithm to blastocyst images to predict a clinical outcome were included. Articles with the following characteristics were excluded: (1) segmentation performed on another form of microscopy than light microscopy or its contrast enhanced variants; (2) descriptions of segmentation algorithms without results; (3) studies on non-human embryos; (4) publications other than journal or conference papers. The full texts of the articles that were of interest were retrieved, and those that met the aforementioned criteria were included for data extraction.

3.3 DATA EXTRACTION

Data were extracted from the selected studies. The primary outcomes collected from the studies were the segmentation evaluation outcome metrics (accuracy, precision, recall, Jaccard Index and Sørensen-Dice Score). Additionally, date of publication, type of images (time-lapse or static acquisition), dataset, potential clinical outcome and the used algorithm were collected. To compare the segmentation performance across the included papers, the Jaccard index was obtained for all segmentation studies. The Jaccard index is often used in segmentation tasks (Eelbode et al. 2020) and measures the similarity between a predicted segmentation mask and the ground truth (i.e. a manual delineation of the blastocyst by an expert embryologist). A Jaccard index of 100% denotes perfect similarity. For studies that did not report Jaccard index as segmentation outcome measure, we calculated this ourselves, either through the reported values for precision (positive predictive value) and recall (sensitivity), or using the Dice score, as presented in formulae 1 and 2.

$$\text{Jaccard Index} = \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall} - (\text{precision} \cdot \text{recall})} \quad (1)$$

$$\text{Jaccard Index} = \frac{\text{Dice Score}}{2 - \text{Dice Score}} \quad (2)$$

3.4 Risk of bias assessment

The risk of bias was assessed for each of the included papers using the Quality Assessment Tool for Diagnostic Accuracy Studies-2 (QUADAS-2) (Whiting et al. 2011). This tool contains four domains: patient selection, index test, reference standard and flow and timing. Domains are assessed as having low, unclear or high risk of bias. If no outcome other than a segmentation algorithm was described, the segmentation algorithm itself was considered the index test.

4 Results

4.1 Study identification

The literature search yielded 417 studies: 148 from Web of Science, 141 from EMBASE, 69 from Medline, 50 from Google Scholar and 9 from Cochrane. After initial removal of 141 duplicate articles, 276 records (89 from Web of Science, 83 from EMBASE, 69 from Medline, 29 from Google Scholar and 6 from Cochrane) were screened on title and abstracts using Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org). After screening, 240 articles were excluded. For the remaining studies, full texts of articles were acquired. For one study, the full text was not available. After screening of the full texts, 22 articles were included. The included articles were published from 2012 to 2023. The screening and inclusion process is summarized in a PRISMA flow diagram (Figure 1).

4.2 Characteristics of included studies

Table 1 displays the basic characteristics of the 22 included studies. Three main types of articles were identified: 17 out of 22 articles (Singh et al. 2015, Kheradmand et al. 2016, Kheradmand et al. 2017, Rad et al. 2017, Saeedi et al. 2017, Rad et al. 2018a, Rad et al. 2018b, Harun et al. 2019a, Harun et al. 2019b, Rad et al. 2019b, Rad et al. 2020, Hu et al. 2021, Arsalan et al. 2022a, Arsalan et al. 2022b, Mushtaq et al. 2022, Farias et al. 2023, Ishaq et al. 2023) only described the development and validation of a segmentation algorithm aimed at one or multiple components of the blastocyst. One article described the development and validation performance of a segmentation algorithm, and additionally investigated the association of their segmentation results to the Gardner grade (Santos Filho et al. 2012). Two articles correlated their developed segmentation algorithm results to blastocyst implantation (Rad et al. 2019a, Fruchter-Goldmeier et al. 2023). Finally, two studies employed an already-developed segmentation tool (Harun et al. 2019b) to predict implantation (Huang et al. 2021) and blastocyst euploidy (Hori et al. 2023).

4.3 Segmentation algorithms and blastocyst components

Three main approaches to segmentation of the blastocyst components were used in the included papers. Four articles employed traditional image processing methods: level-set algorithms and texture analysis (Santos Filho et al. 2012, Singh et al. 2015, Rad et al. 2017, Saeedi et al. 2017). Two studies used manual feature extraction in combination with an artificial neural network (ANN) (Kheradmand et al. 2016, Farias et al. 2023). Finally, the majority of included papers adopted a deep learning approach using a convolutional neural network (CNN) (Kheradmand et al. 2017, Rad et al. 2018a, Rad et al. 2018b, Harun et al. 2019a, Harun et al. 2019b, Rad et al. 2019a, Rad et al. 2019b, Rad et al. 2020, Hu et al. 2021, Huang et al. 2021, Arsalan et al. 2022a, Arsalan et al. 2022b, Mushtaq et al. 2022,

Fruchter-Goldmeier et al. 2023, Hori et al. 2023, Ishaq et al. 2023). While the general structure of CNN models are similar, the specific architecture varies. The architecture defines the specific organisation of the convolutional layers and the flow of mathematical operations within a model. Various CNN architectures were used: Fully Convolutional Neural Networks (FCNN) such as VGG-16, U-Net variations (inception, residual and dilated U-Nets), DeepLabV3 with atrous convolutions, architectures incorporating multi-scale , aggregation, Inception modules, and Scale-Attention networks.

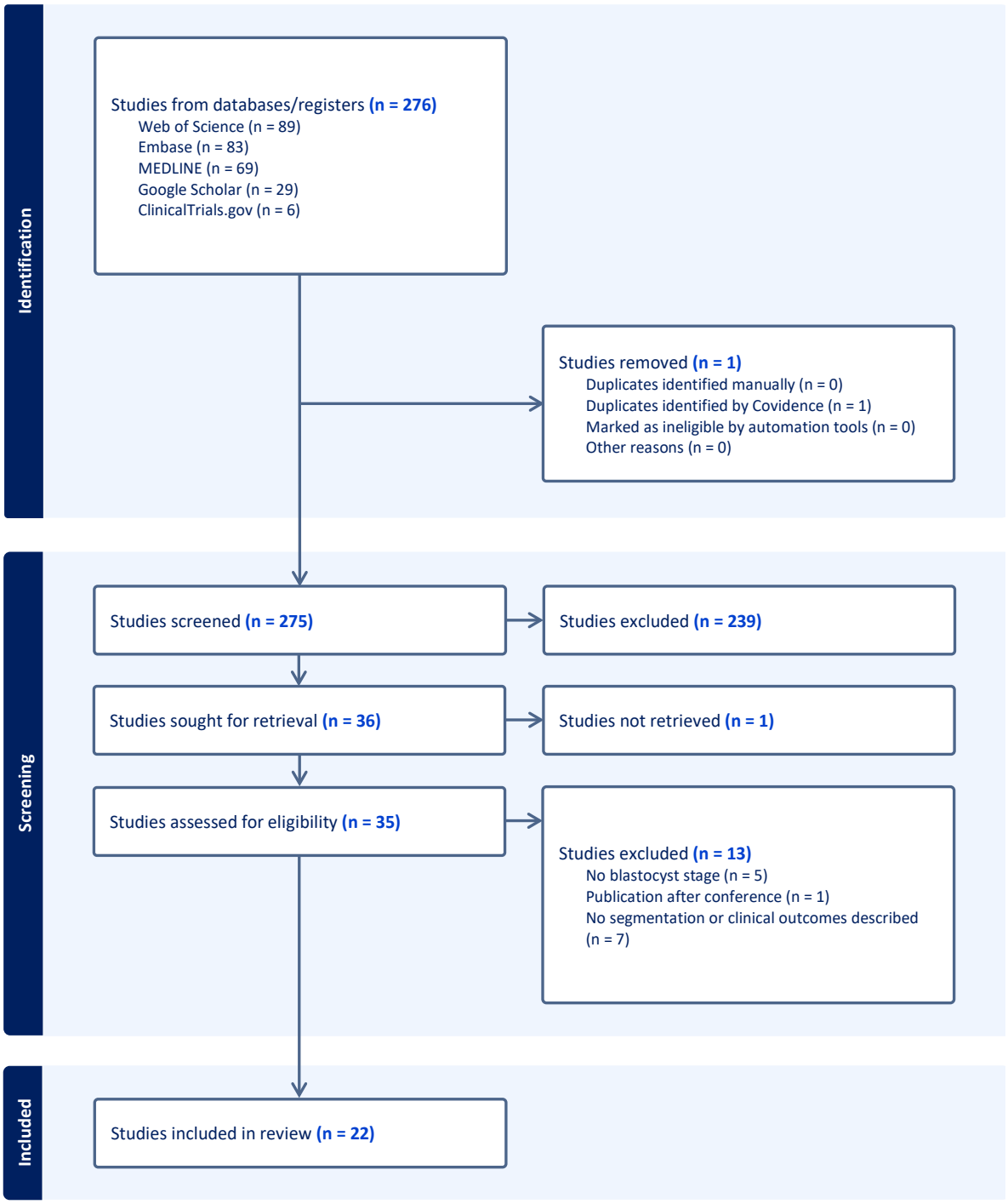


Figure 1: PRISMA flowchart of included studies.

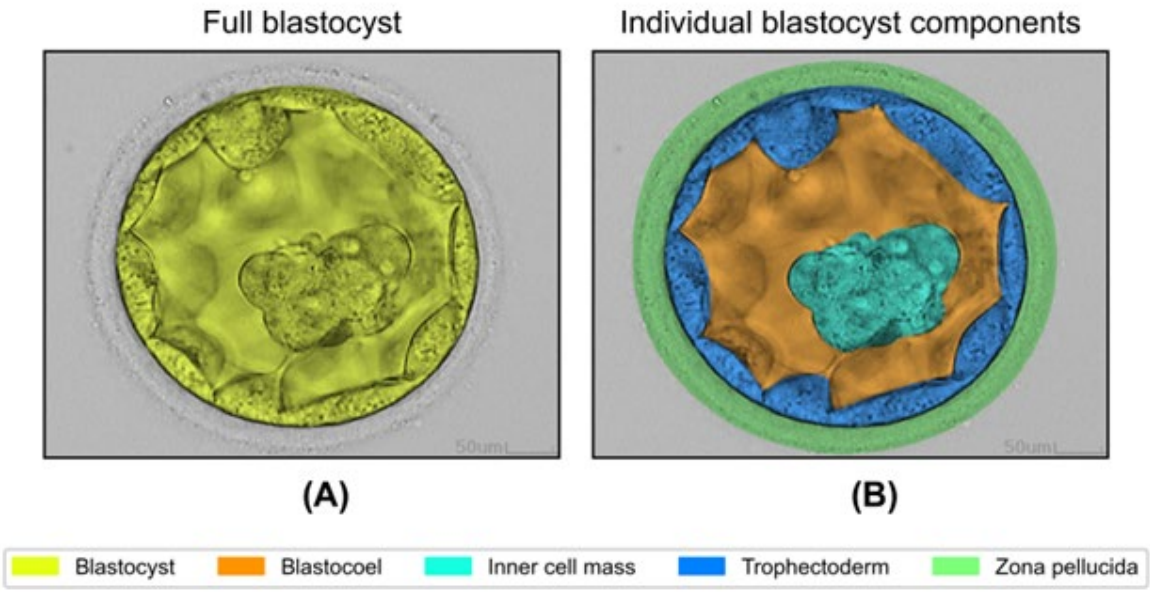


Figure 2: Blastocyst components studied by the included articles. A) segmentation of the blastocyst excluding the zona pellucida. B) the blastocyst with its individual components (blastocoel, inner cell mass, trophectoderm and zona pellucida) highlighted. The blastocyst image is provided in the public

Several components of the blastocyst were segmented and an overview of the different components studied per paper included was generated (Figure 2). Two articles proposed a CNN for segmentation of the blastocyst excluding the ZP (Figure 2A) (Harun et al. 2019b, Fruchter-Goldmeier et al. 2023). In eleven papers, a subset of blastocyst components (ICM, TE, ZP) was segmented, either through a multiclass algorithm or distinct image processing methods (Santos Filho et al. 2012, Singh et al. 2015, Kheradmand et al. 2016, Kheradmand et al. 2017, Rad et al. 2017, Saeedi et al. 2017, Rad et al. 2018a, Rad et al. 2018b, Harun et al. 2019a, Rad et al. 2020, Fruchter-Goldmeier et al. 2023). The ICM and TE, currently considered for blastocyst quality in the Gardner and Schoolcraft embryo grading system, were segmented most often (Table 1). Segmentation of ICM involved texture features (Saeedi et al. 2017), features from discrete cosine transformed (DCS) images as input to an ANN (Kheradmand et al. 2016), and CNNs (Rad et al. 2018b, Harun et al. 2019a, Fruchter-Goldmeier et al. 2023). TE was segmented using a level-set algorithm (Singh et al. 2015, Saeedi et al. 2017), features from DCS images (Kheradmand et al. 2016) and using CNNs (Singh et al. 2015, Harun et al. 2019b, Rad et al. 2020). The ZP was segmented using the features from DCS image (Kheradmand et al. 2016) and a CNN (Rad et al. 2018a). Finally, seven studies developed an algorithm capable of simultaneously segmenting BC, ICM, TE, ZP (i.e. all components shown in Figure 2B) and the image background using one classifier

(Rad et al. 2019a, Rad et al. 2019b, Hu et al. 2021, Arsalan et al. 2022a, Arsalan et al. 2022b, Mushtaq et al. 2022, Ishaq et al. 2023).

Table 1: **Characteristics of the 22 included studies.** ZP: zona pellucida, BC: blastocoel, ICM: inner cell mass, TE: trophectoderm, CNN: convolutional neural network, ANN: artificial neural network, DSC: Dice coefficient, PPV: positive predictive value, acc: accuracy, JI: Jaccard index, HSD: Hausdorff distance, FHB: fetal heart beat, NA: not applicable, NR: not reported.

Publication	Image Type	Blastocyst component	Algorithm	Dataset	N images (Train/Val)	Segmentation Outcome	Clinical Outcome
Santos Filho *et al.* (2012)	Static	ICM, TE, ZP	Level-set and texture	In house	73/73	DSC	Gardner Grade
Singh *et al.* (2015)	Static	TE	Level-set	In house	15/85	Shape accuracy, PPV, recall	-
Kheradmand *et al.* (2016)	Static	ICM, TE, ZP	ANN	Saeedi et al. 2017	130/63	Precision, recall, acc	-
Kheradmand *et al.* (2017)	Static	ICM	CNN (FCNN VGG-16)	Saeedi et al. 2017	188/47	Acc, JI	-
Rad *et al.* (2017)	Static	ICM	Texture features and level-set	Saeedi et al. 2017	220/220	Precision, recall, JI	-
Saeedi *et al.* (2017)	Static	ICM, TE	Texture features and level-set	Saeedi et al. 2017	211/211	Precision, recall, specificity, acc, DSC	-
Rad *et al.* (2018a)	Static	ZP	CNN (Hierarchical NN)	Saeedi et al. 2017	77/158	Precision, recall, acc, JI	-
Rad *et al.* (2018b)	Static	ICM	CNN (U-Net)	Saeedi et al. 2017	200/35	Precision, recall, DSC, JI	-
Harun *et al.* (2019a)	Static	ICM, TE	CNN (Residual and dilated U-Net)	Saeedi et al. 2017	212/37	Precision, recall, acc, DSC, JI	-
Harun *et al.* (2019b)	Time-lapse	Full blastocyst (excluding ZP)	CNN (Residual and dilated U-Net)	In house	462/155	Precision, recall, acc, DSC, JI	-

Rad *et al.* (2019a)	Static	BC, ICM, TE, ZP	CNN (DeepLabV3)	In house	492/86	JI	Implantation
Rad *et al.* (2019b)	Static	BC, ICM, TE, ZP	CNN (Custom)	Saeedi et al. 2017	200/35	JI	-
Rad *et al.* (2020)	Static	ZP	CNN (Inception and dilated U-Net)	In house, Saeedi et al. 2017	592/235	Precision, recall, acc, DSC, JI, HSD	
Hu *et al.* (2021)	Static	BC, ICM, TE, ZP	CNN (Scale-Attention network)	Saeedi et al. 2017	200/35	JI	-
Huang *et al.* (2021)	Time-lapse	Full blastocyst (excluding ZP)	CNN	In house	NA/3030	-	Live birth
Arsalan *et al.* (2022a)	Static	BC, ICM, TE, ZP	CNN (Multi-scale aggregation)	Saeedi et al. 2017	200/35	JI	-
Arsalan *et al.* (2022b)	Static	BC, ICM, TE, ZP	CNN (Sprint Semantic Segmentation)	Saeedi et al. 2017	200/35	JI	-
Mushtaq *et al.* (2022)	Static	BC, ICM, TE, ZP	CNN (ECS-Net)	Saeedi et al. 2017	200/35	JI	-
Farias *et al.* (2023)	Static	BC, ICM, TE, ZP	ANN	Multicentre (2 clinics), Saeedi et al.	592/55	DSC	-
Fruchter-Goldmeier *et al.* (2023)	Time-lapse	Full blastocyst (excluding ZP), ICM	CNN (Mask-RCNN)	Multicentre (3 clinics)	NA/NR	JI	Implantation
Hori *et al.* (2023)	Time-lapse	Full blastocyst (excluding ZP)	CNN	In house	NA/215	-	Euploidy
Ishaq *et al.* (2023)	Static	BC, ICM, TE, ZP	CNN (FSBS-Net)	Saeedi et al. 2017	200/35	Precision, recall, DSC, JI	-

4.4 DATASETS

The publicly available dataset presented by Saeedi et al. (2017) was used in 15 studies for the development and validation of the presented methods. This dataset contained randomly selected blastocysts, imaged between 2012 and 2016 using an Olympus IX71 inverted microscope with Nomarski (differential interference contrast) optics. The images were captured on the fifth day after fertilization at 1.6× and a lens objective of 20×. Two expert embryologists provided pixel-level masks for the ICM, TE and ZP. Two papers (Rad et al. 2020, Farias et al. 2023) used a separate dataset for development and validated the method on the dataset of Saeedi et al. (2017). Seven papers (Santos Filho et al. 2012, Singh et al. 2015, Saeedi et al. 2017, Harun et al. 2019b, Rad et al. 2019a, Huang et al. 2021, Hori et al. 2023) used an in-house dataset for development and validation. Finally, Farias et al. (2023) and Fruchter-Goldmeier et al. (2023) employed a multicentre dataset for the training and validation of their methods. Three papers (Huang et al. 2021, Fruchter-Goldmeier et al. 2023, Hori et al. 2023) employed time-lapse images of developing blastocysts, all other studies used static images. Nine of the included articles expanded the dataset size by using image augmentation such as rotation (Kheradmand et al. 2017, Rad et al. 2018a, Harun et al. 2019b, Rad et al. 2020, Arsalan et al. 2022a, Arsalan et al. 2022b, Mushtaq et al. 2022, Farias et al. 2023), flipping (Rad et al. 2018a, Rad et al. 2020, Arsalan et al. 2022a, Arsalan et al. 2022b, Mushtaq et al. 2022, Farias et al. 2023, Ishaq et al. 2023), cropping (Mushtaq et al. 2022, Ishaq et al. 2023), translations (Arsalan et al. 2022a, Arsalan et al. 2022b, Ishaq et al. 2023), resizing (Mushtaq et al. 2022, Ishaq et al. 2023), zooming in- or out (Rad et al. 2018a, Rad et al. 2020), brightness variation (Farias et al. 2023), shearing (Rad et al. 2020) and the addition of noise (Rad et al. 2018a, Rad et al. 2020). The other papers did not mention artificial expansion of their dataset.

4.5 PERFORMANCE OF SEGMENTATION ALGORITHMS

The performance of the identified segmentation algorithms was analysed by comparing the reported and calculated Jaccard indices of the included papers, displayed in Figure 3. The results are further displayed per study in Table B.1 (Appendix B). To discriminate between the different methodologies used, the papers segmenting the four blastocyst components (BC, ICM, TE and ZP) simultaneously are presented first, followed by the studies that segmented a subset of blastocyst components.

The seven studies that simultaneously segmented all components (the bars with the oblique pattern in Figure 3), all used a version of a CNN. The lowest Jaccard indexes were obtained for segmentation of the TE (range 76.3% - 80.2%). Best performance of a blastocyst component was achieved for the ZP (81.1% - 90.9%). No single approach outperformed the others on all components.

Within the algorithms that segmented one or multiple blastocysts components (the bars without oblique pattern in Figure 3), the CNNs also demonstrated superiority. For ICM segmentation, the Jaccard index ranged from 47.7% to 89.3%. Harun et al. outperformed all other studies by employing a CNN, including those that segmented all blastocyst components (Harun et al. 2019a). The three lowest performing algorithms were those utilizing level-set algorithms, texture features and feature extraction in combination with an ANN for segmentation. Fruchter-Goldmeier et al. reported the segmentation results only categorically: 72.8% of the analysed time-lapse frames had a Jaccard index > 75% (Fruchter-Goldmeier et al. 2023).

Jaccard index values ranged from 50.4% to 85.3% for segmentation of TE. Again, Harun et al. (2019a) outperformed all other approaches on the same dataset. In general, the traditional image processing

approaches (Singh et al. 2015, Saeedi et al. 2017) and ANN (Kheradmand et al. 2016) were again outperformed by CNNs (Harun et al. 2019a, Rad et al. 2020).

For the ZP, the Jaccard index values ranged from 60% to 78.1%. Santos Filho et al (not shown in Figure 3). reported the results categorically. Half of their segmentations yielded a Jaccard index greater than 54%, while a substantial portion (14 out of 44) achieved a Jaccard index smaller than 33% (Santos Filho et al. 2012).

Finally, Harun et al. and Fruchter-Goldmeier et al. segmented the full blastocyst area excluding the ZP (Figure 2A) on time-lapse images of developing blastocysts. Harun et al. reported a Jaccard index of 96.9% (Harun et al. 2019b). Fruchter-Goldmeier et al. only reported the segmentation performance in a categorical manner: 99.9% of the analysed time-lapse frames showed a Jaccard index larger than 75% (Fruchter-Goldmeier et al. 2023).

4.6 Application of segmentation algorithms for prediction of Gardner grade

One study extracted features from their segmentation masks: fractal dimension and thickness of TE and texture features from ICM (Santos Filho et al. 2012). These features were used in support vector machines (SVMs) to predict the embryo's Gardner grade. The authors trained a SVM per subgrade of the grading system in a one-versus-rest strategy. The classifications yielded accuracy values of 67%, 46% and 92% for developmental grades 2, 3 and 4 respectively, 67% and 82% for ICM classification (grades B and C respectively) and 53% and 92% for TE classification (grades B and C respectively).

Application of segmentation algorithms for analysis of blastocyst euploidy

Hori et al. utilized a customized artificial intelligence neural network to assess blastocyst expansion during the first 10 hours of expansion, correlating the findings with biopsied genotypic information. The study revealed that euploid blastocysts expanded significantly faster (P=0.001) and reached a 13.5% larger cross-sectional area to aneuploid blastocysts (P=0.0001) (Hori et al. 2023).

Application of segmentation algorithms for clinical outcomes

Only two studies used the segmentation results to predict implantation potential. (Fruchter-Goldmeier et al. 2023) used the trained segmentation model to extract blastocyst size, ICM size and ICM shape Employing uni- and multivariable logistic regressions, a statistically significant association was observed between blastocyst size and implantation potential (odds ratio (OR) 1.02, 95% confidence interval (CI) 1.01 – 1.03; P<0.001). This association was further accentuated when comparing blastocysts exceeding the mean size with those below it, resulting in an increased odds ratio of 1.74 (95% CI 1.22 – 2.50; P=0.001). Notably, an algorithm combining blastocyst size and maternal age did not outperform an algorithm considering age alone, as indicated by comparable area under the receiver operating characteristic curve values (0.70 versus 0.68; P=1). Rad et al. used the segmentation mask together with the original image as input for a second CNN, aimed at predicting the implantation (Rad et al. 2019a). The positive predictive value (PPV), sensitivity and accuracy were 71.1%, 72.7% and 70.9% respectively. These results outperformed a reported accuracy of 50.7% for an expert embryologist.

In one article, the relationship between automatically quantified blastocyst expansion and pregnancies was investigated (Huang et al. 2021). Results revealed significantly larger areas (20,502 µm2 versus

17,885 µm2; P=0.010) and higher expansion rates in euploid blastocysts that resulted in live births, compared to those that did not.

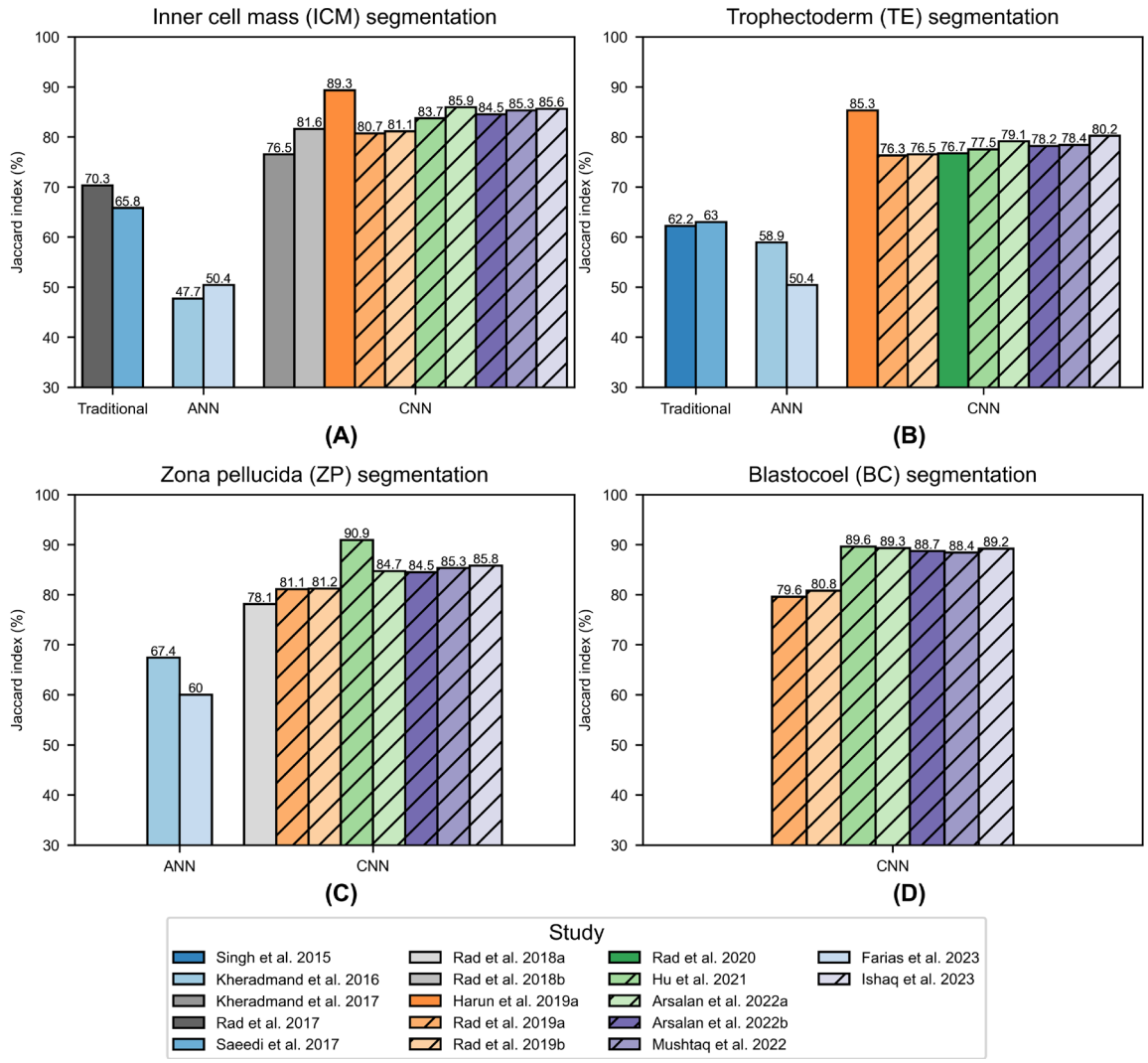


Figure 3: Performance of the blastocyst component segmentation algorithms. Figure A-D display the segmentation performance of the inner cell mass (ICM), trophectoderm (TE), zona pellucida (ZP) and blastocoel (BC) respectively. The different studies are indicated by different colours, while the position on the axis denotes the type of algorithm used: convolutional or artificial neural networks (CNN and ANN), or traditional image processing methods (level-set and texture features). The blastocoel was only segmented using CNNs. Within each method, the studies are ordered on publication year. An oblique pattern denotes that a result was obtained using an algorithm capable of simultaneously segmenting all blastocyst components.

4.7 Risk of bias

The results from the QUADAS-2 assessment are displayed in Figure 4. While few articles display a high risk of bias, a substantial number of included studies had unclear risks regarding the domain index test. It was often not clear how exactly data was used to train models or develop algorithms, or if any threshold optimization was used for the results. The majority of included papers had a low risk of bias regarding patient selection and reference standard. The dataset presented in Saeedi et al. (2017) was utilized in the majority of studies. This dataset was presented with clear in- and exclusion criteria for the images. The individual risk of bias assessments are presented in Table C.1 (Appendix C).

5 Discussion

In this systematic review, a comprehensive overview is provided of the current state of technology pertaining to the automatic segmentation of human blastocyst components. Twenty-two studies were identified, each contributing either to the development or application of a blastocyst segmentation algorithm.

Given that the majority of studies used the dataset by Saeedi et al. (2017), a quantitative comparison of segmentation results could be performed. Harun et al. segmented both the ICM and TE using a CNN, outperforming all other included studies with a Jaccard index of 89.3% and 85.3% respectively (Harun et al. 2019a). Hu et al. demonstrated best performance in BC and ZP segmentation (Jaccard index 89.6% and 90.9%, respectively) using a CNN capable of segmenting all blastocyst components (Hu et al. 2021). These results indicate the feasibility of accurate segmentation of the components of day-5 human blastocysts. CNNs outperformed other image processing methods, such as level-set algorithms, (texture) features and ANNs. The superiority of CNNs can be attributed to their ability to learn directly from images, without the need of manual feature extraction or engineering. This is particularly

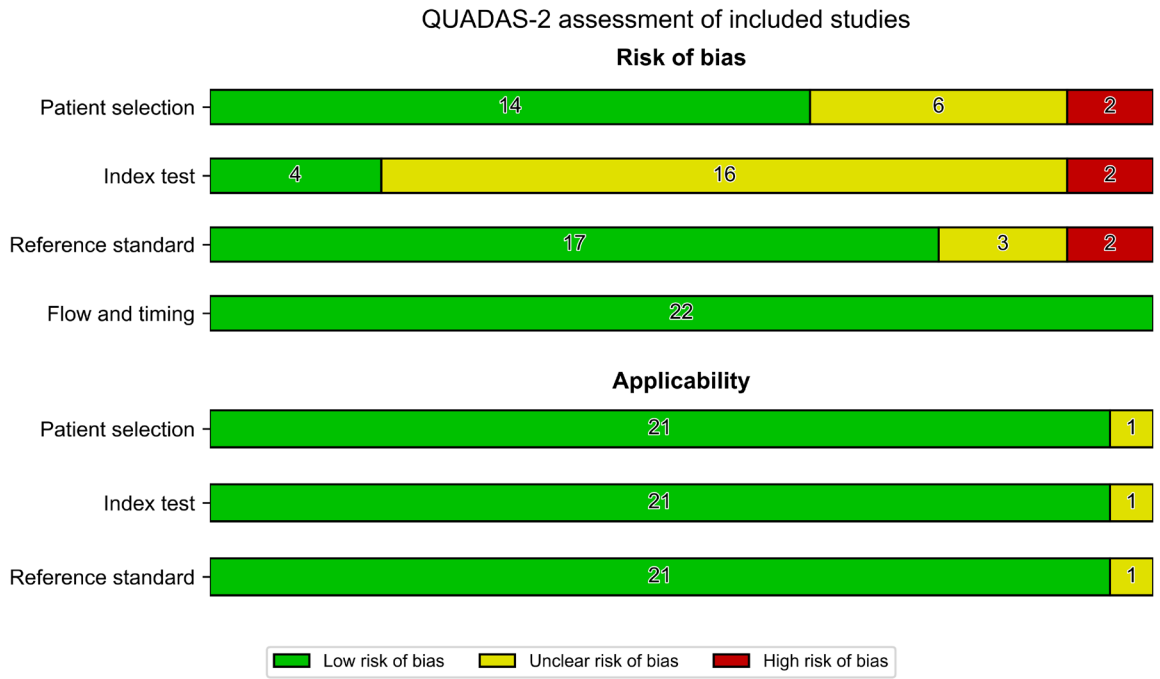


Figure 4: QUADAS-2 risk of bias assessment for the included studies.

advantageous for structures like the ICM and TE, which are of irregular shape and have similar textures. The CNN architectures included in this review aim to learn these complex structures by leveraging multiple image scales and enlarging the receptive field, enabling the model to capture the complexity of these structures. The full areas of developing blastocysts were also segmented with high Jaccard indices (Harun et al. 2019b, Fruchter-Goldmeier et al. 2023). For dynamic images, however, the segmentation results for the individual components were lower than that of static day-5 blastocyst images, indicating the challenges of segmenting blastocyst components at different stages of development (Farias et al. 2023). Developing blastocysts may have more complex and variable shapes, textures and sizes. Employment of a CNN, instead of the ANN used, may improve performance of segmentation of blastocysts from start of blastulation to full expansion and is recommended in future research.

Using proper selection, higher birth rates can be achieved using the transfer of a single blastocyst. Potentially, features extracted from segmentation masks could be used to aid the embryologist in their decision. At this time however, the clinical implications of the segmentation algorithms are not well established in the included papers. One study applied segmentation masks to predict Gardner grades (Santos Filho et al. 2012). However, the subjective nature of Gardner grades raises questions about their clinical utility (Storr et al. 2017) and limits their potential as the gold standard for the segmentation masks. Only three studies have used the segmentation masks to derive quantitative parameters to predict clinical outcomes such as implantation or live birth. Rad et al. used segmentation masks with static images to predict implantation, obtaining promising PPV and sensitivity values (Rad et al. 2019a). Using a segmentation algorithm on time-lapse images, the relationship between automated assessment of blastocyst expansion and live births and euploidy was further investigated. Significant differences were observed in blastocyst expansion between blastocysts that did or did not lead to a pregnancy (Huang et al. 2021). These findings suggest that blastocyst expansion is a potential marker of blastocyst quality and viability, in line with previous findings of manually extracted measurements (Huang et al. 2019, van Marion et al. 2022). The difference in growth rate and cross-sectional area between euploid and aneuploid blastocysts (Hori et al. 2023) might be used to non-invasively gain information about potential chromosomal abnormalities in the blastocyst. Next to the blastocyst expansion, automatic segmentation algorithms also enable analysis of other developmental characteristics, such as the spontaneous implosion of blastocysts during development, known as collapse, which appears to negatively affect the chance of implantation (Cimadomo et al. 2022, Cimadomo et al. 2023). Segmentation algorithms could therefore also be employed for the automatic analysis of large populations of blastocysts, aimed at identifying predictive variables in the full developmental process for implantation and live births.

A limitation of the data presented here, is the lack of external validation reported. Only three studies used multicentre data, while the majority used the dataset by Saeedi et al. (2017) for both the development and validation of the algorithms. The most recent version of this dataset contains 249 images, which is a relatively low number, especially for training CNNs. The relatively small size of this dataset, even with augmentation techniques, may restrict the generalizability of the findings. Addressing this limitation requires validation on datasets from different laboratories, capturing diverse biological variations. Furthermore, future research should focus on expanding datasets to enhance the robustness and applicability of segmentation methods. Finally, the generalizability of the findings might be hampered by the relatively large proportion of conference proceedings. This is reflected in the high number of papers with unsure risk of bias in the domain index test (Figure 4). While these papers have contributed valuable insights, they often lack the stringent peer-review process associated

with journal publications. Additionally, the observed concentration of contributions from a limited number of research groups may have introduced bias.

This systematic review evaluated the current state of technology regarding segmentation techniques applied to human blastocyst components and their possible clinical applications. The findings underscore the feasibility of accurate segmentation of the entire blastocyst during the developmental process and its individual components on the fifth day of development. Convolutional neural networks were found to be the most promising tools for this task and their use should be recommended to researchers in this field. Future research should focus on segmenting individual blastocyst components during the full development process and validating the algorithms on larger and more diverse datasets. Most importantly, the focus in the field can now shift from the development of segmentation tools to the utilisation of these tools for embryo selection and the prediction of clinical outcomes.

6 ACKNOWLEDGEMENTS

The authors wish to thank C. Niehot from the Erasmus MC Medical Library for developing and updating the search strategies.

7 CONFLICTS OF INTERESTS

The authors have no conflicts of interest to disclose.

8 REFERENCES

Alpha Scientists in Reproductive, M., Embryology, E.S.I.G., 2011. Istanbul consensus workshop on embryo assessment: Proceedings of an expert meeting. Reproductive Biomedicine Online. 22, 632-646

Arsalan, M., et al., 2022a. Human blastocyst components detection using multiscale aggregation semantic segmentation network for embryonic analysis. Biomedicines. 10,

Arsalan, M., et al., 2022b. Detecting blastocyst components by artificial intelligence for human embryological analysis to improve success rate of in vitro fertilization. J Pers Med. 12,

Bouillon, C., et al., 2017. Obstetric and perinatal outcomes of singletons after single blastocyst transfer: Is there any difference according to blastocyst morphology? Reproductive Biomedicine Online. 35, 197-207

Cimadomo, D., et al., 2022. Human blastocyst spontaneous collapse is associated with worse morphological quality and higher degeneration and aneuploidy rates: A comprehensive analysis standardized through artificial intelligence. Human Reproduction. 37, 2291-2306

Cimadomo, D., et al., 2023. Opening the black box: Why do euploid blastocysts fail to implant? A systematic review and meta-analysis. Hum Reprod Update. 29, 570-633

Cox, C.M., et al., 2022. Infertility prevalence and the methods of estimation from 1990 to 2021: A systematic review and meta-analysis. Hum Reprod Open. 2022, hoac051

Eelbode, T., et al., 2020. Optimization for medical image segmentation: Theory and practice when evaluating with dice score or jaccard index. Ieee Transactions on Medical Imaging. 39, 3679-3690

Farias, A.F., et al., 2023. Automated identification of blastocyst regions at different development stages. Sci. rep. 13, 15

Fruchter-Goldmeier, Y., et al., 2023. An artificial intelligence algorithm for automated blastocyst morphometric parameters demonstrates a positive association with implantation potential. Sci. rep. 13, 14617

Gardner, D.K., 1999. Towards reproductive certainty: Infertility and genetics beyond 1999. (No Title). 378

Gardner, D.K., et al., 2000. Blastocyst score affects implantation and pregnancy outcome: Towards a single blastocyst transfer. Fertil Steril. 73, 1155-8

Harun, M.Y., et al., 2019a. Inner cell mass and trophectoderm segmentation in human blastocyst images using deep neural network. 13th IEEE International Conference on Nano/Molecular Medicine & Engineering. 214-219

Harun, M.Y., et al., 2019b. Image segmentation of zona-ablated human blastocysts. 13th IEEE International Conference on Nano/Molecular Medicine & Engineering. 208-213

Hori, K., et al., 2023. Comparison of euploid blastocyst expansion with subgroups of single chromosome, multiple chromosome, and segmental aneuploids using an ai platform from donor egg embryos. J Assist Reprod Genet. 40, 1407-1416

Hu, J., et al., 2021. Sa-net: A scale-attention network for medical image segmentation. PLoS ONE. 16, e0247388

Huang, T.T., et al., 2019. Early blastocyst expansion in euploid and aneuploid human embryos: Evidence for a non-invasive and quantitative marker for embryo selection. Reprod Biomed Online. 39, 27-39

Huang, T.T.F., et al., 2021. Deep learning neural network analysis of human blastocyst expansion from time-lapse image files. Reprod Biomed Online. 42, 1075-1085

Ishaq, M., et al., 2023. Assisting the human embryo viability assessment by deep learning for in vitro fertilization. Mathematics. 11,

Kheradmand, S., et al., 2016. Human blastocyst segmentation using neural network. 2016 Ieee Canadian Conference on Electrical and Computer Engineering.

Kheradmand, S., et al., 2017. Inner cell mass segmentation in human hmc embryo images using fully convolutional network. 2017 IEEE International Conference on Image Processing.

Lagalla, C., et al., 2015. A quantitative approach to blastocyst quality evaluation: Morphometric analysis and related ivf outcomes. J Assist Reprod Genet. 32, 705-12

Mushtaq, A., et al., 2022. Artificial intelligence-based detection of human embryo components for assisted reproduction by in vitro fertilization. Sensors (Basel). 22,

Oron, G., et al., 2014. The association between embryo quality and perinatal outcome of singletons born after single embryo transfers: A pilot study. Human Reproduction. 29, 1444-1451

Page, M.J., et al., 2021. The prisma 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 372, n71

Rad, R.M., et al., 2019a. Predicting human embryos' implantation outcome from a single blastocyst image. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Rad, R.M., et al., 2017. Coarse-to-fine texture analysis for inner cell mass identification in human blastocyst microscopic images. Proceedings of the 2017 Seventh International Conference on Image Processing Theory, Tools and Applications.

Rad, R.M., et al., 2018a. Human blastocyst's zona pellucida segmentation via boosting ensemble of complementary learning. Inform Med Unlocked. 13, 112-121

Rad, R.M., et al., 2018b. Multi-resolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images. 2018 25th Ieee International Conference on Image Processing. 3518-3522

Rad, R.M., et al., 2019b. Blast-net: Semantic segmentation of human blastocyst components via cascaded atrous pyramid and dense progressive upsampling. 2019 Ieee International Conference on Image Processing. 1865-1869

Rad, R.M., et al., 2020. Trophectoderm segmentation in human embryo images via inceptioned u-net. Med Image Anal. 62, 101612

Saeedi, P., et al., 2017. Automatic identification of human blastocyst components via texture. IEEE Trans Biomed Eng. 64, 2968-2978

Santos Filho, E., et al., 2012. A method for semi-automatic grading of human blastocyst microscope images. Hum Reprod. 27, 2641-2648

Singh, A., et al., 2015. Automatic segmentation of trophectoderm in microscopic images of human blastocysts. IEEE Trans Biomed Eng. 62, 382-393

Storr, A., et al., 2017. Inter-observer and intra-observer agreement between embryologists during selection of a single day 5 embryo for transfer: A multicenter study. Human Reproduction. 32, 307-314

Thompson, S.M., et al., 2013. Blastocyst expansion score and trophectoderm morphology strongly predict successful clinical pregnancy and live birth following elective single embryo blastocyst transfer (eset): A national study. J Assist Reprod Genet. 30, 1577-81

Van Marion, E.S., et al., 2022. Longitudinal surface measurements of human blastocysts show that the dynamics of blastocoel expansion are associated with fertilization method and ongoing pregnancy. Reproductive Biology and Endocrinology. 20, 53

Whiting, P.F., et al., 2011. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 155, 529-36

APPENDIX A

Table A.1: Overview of searched databases and records identified.

Database searched	Platform	Years of coverage	Records	Records after duplicates removed
Medline ALL	Ovid	1946 - Present	69	69
Embase	Embase.com	1971 - Present	141	83
Web of Science Core Collection*	Web of Knowledge	1975 - Present	148	89
Cochrane Central Register of Controlled Trials**	Wiley	1992 - Present	9	6
Additional Search Engines: Google Scholar***			50	29
Total			**417**	**276**

*Science Citation Index Expanded (1975-present) ; Social Sciences Citation Index (1975-present) ; Arts & Humanities Citation Index (1975-present) ; Conference Proceedings Citation Index- Science (1990-present) ; Conference Proceedings Citation Index- Social Science & Humanities (1990-present) ; Emerging Sources Citation Index (2005-present)
*Exact search turned on in Web of Science Core Collection
** Manually deleted abstracts from trial registries
***Google Scholar was searched via "Publish or Perish" to download the results in EndNote.

No other database limits were used than those specified in the search strategies

Medline

((segment*).ab,ti,kf.) **AND** ("Automation"/ OR exp "Artificial Intelligence"/ OR "Algorithms"/ OR (automat* OR AI OR artificial-intell* OR neural-network* OR ((machine* OR deep*) ADJ3 (learn*)) OR algorhythm* OR algorithm*).ab,ti,kf.) **AND** (exp "Blastocyst"/ OR *"Embryonic Structures"/ OR *" Embryo"/ OR *"Embryonic Development"/ OR (blastocyst* OR blastocoel* OR embryoblast* OR inner-cell-mass* OR trophectoderm*).ab,ti,kf. OR (embryo*).ti.) **NOT** (exp animals/ NOT humans/)

Embase

('embryo segmentation'/de/mj OR 'segmentation algorithm'/exp OR (segment*):ab,ti,kw) **AND** ('automation'/de OR 'artificial intelligence'/de OR 'machine learning'/de OR 'feature detection'/exp OR 'feature extraction'/exp OR 'feature learning (machine learning)'/exp OR 'algorithm'/de OR 'artificial neural network'/exp OR (automat* OR AI OR artificial-intell* OR neural-network* OR ((machine* OR deep*) NEAR/3 (learn*)) OR algorhythm* OR algorithm*):ab,ti,kw) **AND** ('blastocyst'/exp OR 'embryo segmentation'/de/mj OR 'embryo development'/de/mj OR 'embryo (anatomy)'/de/mj OR (blastocyst* OR blastocoel* OR embryoblast* OR inner-cell-mass* OR trophectoderm*):ab,ti,kw OR (embryo*):ti) NOT ((animal/exp OR animal*:de OR nonhuman/de) NOT ('human'/exp))

Web of Science

TS=(segment*) **AND** TS=(automat* OR AI OR artificial-intell* OR neural-network* OR ((machine* OR deep*) NEAR/2 (learn*)) OR algorhythm* OR algorithm*) **AND** (TS=(blastocyst* OR blastocoel* OR embryoblast* OR inner-cell-mass* OR trophectoderm*) OR TI=(embryo*)) **NOT** TS=((animal* OR rat

OR rats OR mouse OR mice OR murine OR dog OR dogs OR canine OR cat OR cats OR feline OR rabbit OR cow OR cows OR bovine OR rodent* OR sheep OR ovine OR pig OR swine OR porcine OR veterinar* OR chick* OR zebrafish* OR baboon* OR nonhuman* OR primate* OR cattle* OR goose OR geese OR duck OR macaque* OR avian* OR bird* OR fish*) NOT (human* OR patient* OR women OR woman OR men OR man))

Cochrane CENTRAL

((segment*):ab,ti,kw) **AND** ((automat* OR AI OR artificial NEXT/1 intell* OR neural NEXT/1 network* OR ((machine* OR deep*) NEAR/3 (learn*)) OR algorhythm* OR algorithm*):ab,ti,kw) **AND** ((blastocyst* OR blastocoel* OR embryoblast* OR inner NEXT/1 cell NEXT/1 mass* OR trophectoderm*):ab,ti,kw OR (embryo*):ti)

Google Scholar

segmentation automated|'artificial intelligence'|AI|'deep|machine learning'|'neural network'|algorithm blastocyst

APPENDIX B

Table B.1: Jaccard index (%) of the studies reporting their segmentation outcomes. Results are shown per blastocyst component: blastocoel (BC), inner cellular mass (ICM), trophectoderm (TE), zona pellucida (ZP) and image background (BG). The parameters column provides the number of trainable parameters used in the neural network. Dash (-): data not reported.

Study	BC	ICM	TE	ZP	BG	Parameters
Singh *et al.* (2015)	-	-	62.2	-	-	-
Kheradmand *et al.* (2016)	-	47.7*	58.9*	67.4*	-	-
Kheradmand *et al.* (2017)	-	76.5*	-	-	-	-
Rad *et al.* (2017)	-	70.3*	-	-	-	-
Saeedi *et al.* (2017)	-	65.8*	63.0*	-	-	-
Rad *et al.* (2018a)	-	-	-	78.1*	-	-
Rad *et al.* (2018b)	-	81.6*	-	-	-	-
Harun *et al.* (2019a)	-	89.3*	85.3*	-	-	-
Rad *et al.* (2019a)	79.6	80.7	76.3	81.1	94.6	-
Rad *et al.* (2019b)	80.8*	81.1*	76.5*	81.2*	94.7	25M
Rad *et al.* (2020)	-	-	76.7*	-	-	-
Hu *et al.* (2021)	89.6*	83.7*	77.5*	90.9*	97.5*	-
Arsalan *et al.* (2022a)	89.3*	85.9*	79.1*	84.7*	96.1*	2.06M
Arsalan *et al.* (2022b)	88.7*	84.5*	78.2*	84.5*	95.8*	4.04M
Mushtaq *et al.* (2022)	88.4*	85.3*	78.4*	85.3*	94.9*	2.83M
Farias *et al.* (2023)	73.9	37.0/50.4*	46.0/50.4*	55.0/60.0*	92.3	-
Ishaq *et al.* (2023)	89.2*	85.6*	80.2*	85.8*	95.6*	2.01 M

**Results with asterisk are obtained using (part of) the dataset presented in (Saeedi et al. 2017).*

APPENDIX C

Table C.1: QUADAS-2 risk of bias assessment reported per study.

	Risk of bias				Applicability		
	Patient selection	**Index test**	**Reference standard**	**Flow and timing**	**Patient selection**	**Index test**	**Reference standard**
Santos Filho *et al.* (2012)	Unclear	Unclear	Low	Low	Low	Low	Low
Singh *et al.* (2015)	High	Unclear	Low	Low	Low	Low	Low
Kheradmand *et al.* (2016)	Unclear	Unclear	Unclear	Low	Low	Low	Low
Kheradmand *et al.* (2017)	Low	Unclear	High	Low	Low	Low	Low
Rad *et al.* (2017)	Unclear	High	Low	Low	Low	Low	Low
Saeedi *et al.* (2017)	Low	High	Low	Low	Low	Low	Low
Rad *et al.* (2018a)	Low	Low	Low	Low	Low	Low	Low
Rad *et al.* (2018b)	Low	Unclear	Low	Low	Low	Low	Low
Harun *et al.* (2019b)	Unclear	Unclear	Low	Low	Low	Low	Low
Harun *et al.* (2019a)	Low	Unclear	Low	Low	Low	Low	Low
Rad *et al.* (2019a)	Unclear	Unclear	High	Low	Low	Low	Unclear
Rad *et al.* (2019b)	Low	Unclear	Unclear	Low	Low	Low	Low
Rad *et al.* (2020)	Unclear	Unclear	Unclear	Low	Low	Low	Low
Hu *et al.* (2021)	Low	Low	Low	Low	Low	Low	Low
Huang *et al.* (2021)	High	Low	Low	Low	Unclear	Unclear	Low
Arsalan *et al.* (2022a)	Low	Unclear	Low	Low	Low	Low	Low
Arsalan *et al.* (2022b)	Low	Unclear	Low	Low	Low	Low	Low

Mushtaq *et al.* (2022)	Low	Unclear	Low	Low	Low	Low	Low
Farias *et al.* (2023)	Low	Unclear	Low	Low	Low	Low	Low
Fruchter-Goldmeier *et al.* (2023)	Low	Unclear	Low	Low	Low	Low	Low
Ishaq *et al.* (2023)	Low	Unclear	Low	Low	Low	Low	Low
Hori *et al.* (2023)	Low	Low	Low	Low	Low	Low	Low