Artificial Intelligence (AI) for Embryo Ranking and its Use in Human Assisted Reproduction

A thesis submitted to the University of Kent for the degree of

DOCTOR OF PHILOSOPHY

in the Division of Natural Sciences

Alejandro Chavez Badiola

November 2023

The School of Biosciences

Declaration

No part of this thesis has been submitted in support of an application for any degree or qualification from the University of Kent.

Dr. Alejandro Chavez-Badiola is a shareholder in two companies related to reproductive medicine and technology. IVF 2.0 LTD UK develops and markets ERICA and other AI technologies for reproductive medicine applications. Dr. Chavez-Badiola also co-founded and is a shareholder in Conceivable Life Sciences, a company created to make reproductive medicine more accessible through automation and AI innovations for IVF procedures. Conceivable Life Sciences owns the intellectual property rights for ERICA.

The software containing the described algorithm and ERICA are registered for trademark and author rights. Intellectual property in the form of patents has been granted in Mexico and the USA and is currently undergoing national phases in Europe and other regions.

Alejandro Chavez-Badiola

November 2023

Incorporation of Published Work

The following thesis incorporates text from the following list of manuscripts. Where I wrote the original text, it appears largely as written, with some changes for context and to fit the style of a thesis. Grammarly and the AI algorithms Claude and ChatGPT have assisted in aid re-writing to improve and standardise the language.

Aspects of the following review are incorporated into Chapter I.

Dimitriadis I, Zaninovic N, Chavez Badiola A, Bormann CL. Artificial intelligence in the embryology laboratory: a review. Reproductive biomedicine online. 2022 Mar 1;44(3):435-48.

Results from Chapter III.

Chavez-Badiola A, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ, Garcia-Sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. Scientific reports. 2020 Mar 10;10(1):4394.

Results from Chapter IV.

Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reproductive BioMedicine Online. 2020 Oct 1;41(4):585-93.

Results from Chapter V.

Chavez-Badiola A, Flores-Saiffe Farías, Gerardo Mendizabal-Ruiz, Silvestri G, Griffin DK, Valencia-Murillo R, Andrew J. Drakeley AJ, Cohen J. Use of artificial intelligence (AI) embryo selection based on static images to predict first-trimester pregnancy loss: a retrospective pilot study. In preparation for publication

Results from Chapter VI.

Farias AF, Chavez-Badiola A, Mendizabal-Ruiz G, Valencia-Murillo R, Drakeley A, Cohen J, Cardenas-Esparza E. Automated identification of blastocyst regions at different development stages. Scientific Reports. 2023 Jan 2;13(1):15.

Results from Chapter VII

Curchoe CL, Farias AF, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. Fertility and sterility. 2020 Nov 1;114(5):921-6.

The following papers were also published during the course of the thesis and are related but not substantive to the thesis:

1. Abdullah KA, Atazhanova T, Chavez-Badiola A, Shivhare SB. Automation in ART: paving the way for the future of infertility treatment. Reproductive Sciences. 2023; 30(4):1006-16.
2. Glatstein I, Chavez-Badiola A, Curchoe CL. New frontiers in embryo selection. Journal of assisted reproduction and genetics. 2023 Feb;40(2):223-34.
3. Mendizabal-Ruiz G, Chavez-Badiola A, Figueroa IA, Nuño VM, Farias AF, Valencia-Murillo R, Drakeley A, Garcia-Sandoval JP, Cohen J. Computer software (SiD) assisted real-time single sperm selection associated with fertilization and blastocyst formation. Reproductive BioMedicine Online. 2022 Oct 1;45(4):703-11.
4. Moreno I, Garcia-Grau I, Perez-Villaroya D, Gonzalez-Monfort M, Bahçeci M, Barrionuevo MJ, Taguchi S, Puente E, Dimattina M, Lim MW, Meneghini G, Aubuchon M, Leondires M, Izquierdo A, Perez-Olgiati M, Chavez Badiola A, Seethram K, Bau D, Gomez C, Valbuena D, Vilella F, Simon C. Endometrial microbiota composition is associated with reproductive outcome in infertile patients. Microbiome. 2022 Dec;10:1-7.
5. Chavez-Badiola A, Mendizabal-Ruiz G, Flores-Saiffe Farias A, Garcia-Sanchez R, Drakeley AJ. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Human Reproduction. 2020 Feb 29;35(2):482.
6. Curchoe CL, Malmsten J, Bormann C, Shafiee H, Farias AF, Mendizabal G, Chavez-Badiola A, Sigaras A, Alshubbar H, Chambost J, Jacques C. Predictive modelling in reproductive medicine: where will the future of artificial intelligence research take us? Fertility and sterility. 2020 Nov 1;114(5):934-40.
7. Sánchez-González CM, Aguinaga-Ríos M, García-Sánchez R, Sánchez-González D, Guarneros-Valdovinos R, Chávez-Badiola A. A combined in-vitro fertilisation strategy: minimal stimulation IVF, PGT-A and SET. Results from 3 years of experience at two IVF centres in Mexico. Ginecología y obstetricia de México. 2020;88(8):508-16.

8. Recio-López Y, López-Rioja MD, Sánchez-González CM, Iñiguez-Arteaga EL, Salas-Rosa FS, Chávez-Badiola A. Flow cytometry sorter: results in the key performance indicators of an assisted reproductive technology. Ginecología y obstetricia de México. 2019;87(1):6-19.

The following are published abstracts

1. Aguilar I, Chavez-Badiola A, Paredes O, Flores-Saiffe Farías A, Drakeley AJ, Sakkas D, Ocali O, Hernández León M, Mendizabal Ruiz G, Valadez Aguilar A, Sánchez Gonzalez D. P-093 Single sperm morphokinetic variables during ICSI at the time of sperm aspiration into the microneedle. Human Reproduction. 2023, (Vol 1;38, 093-457).
2. Canedo C, Camard J, Silvestri G, Marques M, Serrano-Albal M, Robinson G, Bradu A, Chavez-Badiola A, Podoleanu A, Griffin DK. Application of a Time-Lapse Optical Coherence Tomography (OCT) approach in a pilot study to visualise oocytes and embryos in depth. Human Reproduction,. 2023; 38 (S1) P-295: 653. doi.org/10.1093/humrep/dead093.653
3. Chavez-Badiola, A., Flores-Saiffe Farías, A., Mendizabal, G., Valencia-Murillo, R., Sakkas, D., Ocali, O., Mazur, P., Viñals Gonzalez, X., Hernandez Leon, M., Valadez Aguilar, A. and Griffin, D. Use of an artificial intelligence tool to assess single-sperm motility variables related to bias preference of ICSI sperm selection practice, normal fertilization, and blastocyst formation. Human Reproduction. 2022, (Vol. 37,107-081).
4. Chavez-Badiola A, Flores-Saiffe Farias A, Sanchez D, Mendizabal-Ruiz G, Valencia-Murillo R, Drakeley A, Cohen J. The location of fragments and degraded zones in blastocysts is associated with ploidy: moving towards explaining an AI-based morphology tool trained on euploidy outcomes. Human Reproduction, 2022 (Vol. 37, 107-239).
5. Chavez Badiola, A., Flores-Saiffe, A., Valencia, R., Mendizabal-Ruiz, G., Villavicencio, J., Gonzalez, D., Griffin, D., Drakeley, A. and Cohen, J., 2022. P-241 'Augmented intelligence' to possibly shorten euploid identification time: A human-machine interaction study for euploid identification using ERICA, an Artificial Intelligence software to assist embryo ranking. Human Reproduction, 2022 (Vol. 37,107-231).
6. González DJ, Farías AF, Valadez A, Hernandez MI, Chavez-Badiola A. An Artificial Intelligence Model Can Anticipate Embryo Ploidy Potential, Using Its Score As Predictive Value. Fertility and Sterility. 2022; 1;118(4): e112-3.
7. Valencia R, Farías AF, Mendizabal G, Chavez-Badiola A, Gomez FJ. Towards An Explainable Artificial Intelligence To Predict Blastocyst Formation Potential From Single Oocyte Images. Fertility and Sterility. 2022;118(4):e337-8.

8. Farías AF, Sakkas D, Chavez-Badiola A, Ocali O, Mendizabal G, Valencia R, Valadez A, Hernandez MI, Drakeley AJ, Cohen J. Single-Sperm Motility Analysis During Icsi Using An Artificial Intelligence Sperm Identification Software (SID) And Correlation With Morphology. Fertility and Sterility. 2022;118(4):e56-7.
9. Chavez-Badiola A, Farias A, Mendizabal-Ruiz G, Griffin D, Valencia-Murillo R, Reyes-Gonzalez D, Drakeley AJ, Cohen J. ERICA (Embryo Ranking Intelligent Classification Assistant) AI predicts miscarriage in poorly ranked embryos from one static, non-invasive embryo image assessment. Human Reproduction 2021 (Vol. 36, pp. 128-059.).
10. Chavez-Badiola A, Farias AF, Valencia R, Drakeley AJ, Cohen J. Improving Erica's (Embryo Ranking Intelligent Classification Assistant) Performance. Should We Train An AI To Remain Static Or Dynamic, Adapting To Specific Conditions?. Fertility and Sterility. 2021 Sep 1;116(3):e151-2.
11. Drakeley A, Flores-Saiffe A, Chavez-Badiola A, Reyes-Gonzalez D, Valencia R, Mendizabal-Ruiz G, Cohen J. ERICA embryo AI ranking based on ploidy prediction correlates with pregnancy outcomes. BJOG, 2021 (Vol. 128, pp. 231-232).
12. Hernández MM, Flores DA, Chavez-Badiola A, Mendizabal-Ruiz G. Diseño e implementación de un simulador de inyección intracitoplasmática (ICSI). InMemorias del Congreso Nacional de Ingeniería Biomédica 2021 Nov 30 (Vol. 8, No. 1, pp. 159-163).
13. Pina-Aguilar R, Callul-Bagó A, Aguinaga M, González-Ortega C, Pascual-Rodríguez A, Pérez-Peña E, Iñiguez-Arteaga E, Sánchez-González D, Martínez-Garza S, Chávez-Badiola A, Gutiérrez-Gutiérrez A. Successful experiences of preimplantation genetic testing for monogenic diseases (PGT-M): a step forward for clinical genetics in Mexico. Molecular Genetics and Metabolism. 2021;132:S306-7.
14. Chavez-Badiola A. Flores-Saiffe Farias A, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ, Garcia-Sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. Sci Rep. 2020 Mar 10;10(1):4394.
15. Chavez-Badiola A, Farias AF, Valencia R, Mendizabal-Ruiz G, Drakeley AJ. A Computer-Vision Based Tool For The Automatic Identification Of Blastocysts' Regions. A Step Closer To Decoding Time-Lapse? Fertility and Sterility. 2020;114(3):e142-3.
16. Chavez-Badiola A, Zhang JJ, Farias AF, Olcha M, Drakeley AJ. Non-invasive chromosome screening and its correlation against ranking prediction made by ERICA, a deep-learning embryo ranking algorithm. Fertility and Sterility. 2020;114(3):e436-7.

17.Chavez-Badiola A, Farias AF, Valencia R, Drakeley AJ. Automated Identification of Degraded Areas Within Blastocysts By Means of Artificial Vision. Fertility and Sterility. 2020 Sep 1;114(3):e138.

18.Figueroa IA, Nuño JM, Badiola AC, Ruíz EM. Sistema Basado en Visión Computacional para Asistir en Selección de Espermatozoides en Protocolos de Fertilización in-vitro. In Memorias del Congreso Nacional de Ingeniería Biomédica 2020. 2020; (Vol. 7, No. 1, pp. 69-76).

19.Nuño JM, Figueroa IA, Badiola AC, Ruíz EM. Sistema Inteligente para el Monitoreo Automático del Desarrollo de embriones dentro de Incubadoras. InMemorias del Congreso Nacional de Ingeniería Biomédica 2020. 2020; (Vol. 7, No. 1, pp. 525-531).

20.Chavez-Badiola A, Farias AF, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ. Development and preliminary validation of an automated static digital image analysis system utilizing machine learning for blastocyst selection. Fertility and Sterility. 2019;112(3):e149-50.

21.Chavez-Badiola A, Mendizabal-Ruiz G, Ocegueda-Hernandez V, Farias AF, Drakeley AJ. Deep learning for automatic determination of blastocyst embryo development stage. Fertility and Sterility. 2019 Sep 1;112(3):e273.

22.Chavez-Badiola A, Farias AF, Mendizabal-Ruiz G, Drakeley AJ, Garcia-Sánchez R, Zhang JJ. Artificial vision and machine learning designed to predict PGT-A results. Fertility and Sterility. 2019;112(3):e231.

23.Chavez-Badiola A, Garcia-Sánchez R, Iñiguez-Arteaga EL, González DJ, Madera CO, Vazquez-Pacheco S. Meiotic spindle avoidance using a polarizing filter at the time of ICSI, a step closer towards individualized ICSI. Fertility and Sterility. 2019; 112(3):e432-3.

24.Recio-López Y, de Jesús López-Rioja M, Sánchez-González CM, Iñiguez-Arteaga EL, Salas-Rosa FS, Chávez-Badiola A. Sorter de citometría de flujo: repercusiones en los indicadores clave de rendimiento de un laboratorio de reproducción asistida. Ginecología y obstetricia de México. 2019; 87(01):6-19.

All appear, as published, in the appendix.

Guapa, Ines and Santiago, to you with immense love.

This achievement is ours to share.

We write to taste life twice, in the moment, and in retrospection. We write, like Proust, to render all of it eternal, and to persuade ourselves that it is eternal. We write to be able to transcend our life, to reach beyond it.

Anaïs Nin, entry for February 1954, in The Diary of Anaïs Nin

Acknowledgements

My deepest love and gratitude to my wife Daniela, my rock. My sweet children Ines and Santiago, through long days and longer nights, your hugs invigorate me to strive higher. I could not ask for a more supportive family, nor one I cherish more deeply. You give my ambitions purpose and my accomplishments meaning. With my whole heart, I thank you.

I want to thank my parents and siblings, especially Cristina, whose hard work and dedication have inspired me for many years. To John and the New Hope Fertility Center Mexico teams for their unconditional support during this and every endeavour I've pursued. Thank you.

My admiration for the teams at IVF 2.0 and Conceivable Life Sciences, where my passion for innovation and lifelong learning has been nurtured. Gerardo, Adolfo, and Paulo thank you for patiently introducing me to a whole new world where engineering, mathematics, medicine and magic become one. Joshua and Alan, it's a joy to share a dream and to be able to work towards it alongside you.

I am tremendously grateful to Jacques Cohen for insisting that I undertake this PhD program and to my supervisor, Professor Darren Griffin, for his friendship and mentorship have been invaluable. Peter Ellis, thank you for your help and support, especially over the past months as a pinch hitter.

To the thousands of patients who have entrusted me with their dreams of building families during almost two decades in practice. My sincerest appreciation to everyone who has contributed to this journey.

Table of Contents

1. General Introduction

1.1. Artificial intelligence and its impact on society

1.1.1. AI in finance

1.1.2. AI in industry

1.1.3. AI Techniques in Image Analysis and Processing

1.1.3.1. Deep Learning

1.1.3.2. Computer Vision Techniques

1.1.3.2.1. Feature extraction

1.1.3.2.2. Image segmentation

1.1.3.2.3. Image representation

1.1.3.2.4. Image filtering

1.1.3.2.5. Image understanding and image registration

1.1.4. AI in medicine

1.1.4.1. The introduction of AI models. Its effect on developers and users

1.1.4.2. AI and data sources

1.1.4.3. AI validation

1.1.4.4. Critical appraisal of an AI model

1.1.4.5. Ground truth and validation set

1.2. Infertility, Assisted Reproduction Technology (ART) and In-Vitro Fertilisation (IVF)

1.2.1 Causes of infertility

1.2.2 Treatments for infertility

1.2.3 IVF and ICSI

1.2.4. Morphological analysis of embryos

1.2.4.1 Classic analysis

1.2.4.2 Time-lapse and morphokinetics

1.2.4.3 Image-based embryo evaluation and Artificial Intelligence

1.2.5. Preimplantation genetic testing (PGT)

1.2.5.1. PGT for the diagnosis of genetic disease

1.2.5.2. PGT-A for the improvement of IVF success

1.2.6. The relationship between embryo morphology and aneuploidy

1.3 The need for AI in ART with specific reference to Image-based evaluation

1.3.1. AI and gametes

1.3.1.1. AI and reproductive urology

1.3.1.2. AI and the impact of controlled ovarian stimulation on oocytes

1.3.2. AI and embryo assessment

1.3.2.1. Pronuclear-stage assessment

1.3.2.2. AI and cleavage stage-assessment

1.3.2.3. AI assessment of Blastocyst stage embryos

1.3.2.4. Training databases

1.3.2.5. Time-lapse microscopy (TLM) image analysis

1.3.2.6. Single-image analysis

1.3.2.7. Automatic annotation

1.3.2.8. Implantation prediction

1.3.2.9. AI for detecting aneuploidy

1.4. Thesis perspectives

1.5 Thesis aims

2. Materials and Methods

2.1. Ovarian stimulation and embryo culture

2.2. Embryo transfer

2.3. Ethical approval

3. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning

3.1. Chapter Summary

3.2. Chapter Aims

3.3. Materials and Methods

3.3.1 Positive b-hCG and confirmation

3.3.2 Image analysis and statistical assessment

3.4. Results

3.4.1 Embryo identification capability

3.4.2 Positive b-hCG prediction

3.5. Discussion

4. Embryo Ranking Intelligent Classification Algorithm (ERICA), an artificial intelligence clinical assistant with embryo ploidy and implantation predicting capabilities

4.1. Chapter Summary

4.2. Chapter Introduction

4.3. Chapter Aims

4.4 Materials and Methods

- 4.4.1. Database description.
- 4.4.2. Training and testing.
- 4.4.3. Algorithm description.
- 4.4.4. Evaluating the algorithm.
- 4.4.5. Ethical considerations.

4.5. Results

- 4.5.1. Training ERICA.
- 4.5.2. Testing ERICA.

4.6. Discussion

5. Use of artificial intelligence (AI) embryo selection (ERICA) based on static images to predict first-trimester pregnancy loss: a retrospective pilot study

5.1. Chapter Summary

5.2. Chapter introduction

- 5.2.1. Chapter specific aims

5.4. Materials and Methods

- 5.4.1 Ethical approval, data collection, definitions, and study inclusion criteria
- 5.4.2. Ovarian stimulation and embryo culture
- 5.4.3. Embryo transfer
- 5.4.4 Embryo Imaging and AI-Assisted Ranking
- 5.4.5 Statistical analysis

5.5. Results

- 5.4.1 Relationship between early pregnancy loss (SA), embryo ranking, and recipient age
- 5.4.2. SA in euploid embryos determined by PGT-A

5.6. Discussion

6. Automated identification of blastocyst regions at different development stages

6.1. Chapter Summary

6.2. Chapter introduction

6.2.1 Chapter aims

6.3. Material and Methods

6.3.1. Database and ground truth description

6.3.2. Image pre-processing and model training

6.3.2.1. Image pre-processing

6.3.2.2. Pixel classification

6.3.2.3. Segmentation refinement

6.3.3. Sensitivity analysis and validation

6.4. Results

6.5. Discussion

7. Evaluating AI predictive models in reproductive medicine

7.1 Chapter Summary

7.2. Chapter Introduction

7.3. Material and Methods

7.3.1. Inclusion Criteria

7.3.2. Exclusion Criteria

7.3.3. Study Selection

7.4. Results

7.5. Chapter discussion

7.5.1. Four reasons to interpret reproduction AI studies with caution

7.5.2. Reporting and validating results

7.5.3. Ground truth and validation set

7.5.4. Database features

7.5.5. Reproducibility and repeatability

7.5.6. Sensitivity and specificity testing

7.5.7. Results provided

7.5.8. Separate data for validation and testing

7.5.9. Test set comparison to user settings

7.6. Reader and Referee Considerations: A Quick Reference

7.7. Critical appraisal of AI in reproductive medicine

8. General Discussion

8.1. Promise and Progress of AI in Embryo Assessment

8.2 Bridging the Gap Between Potential and Practice

8.3 Responsible Translation Into the IVF Laboratory

8.4 The Road Ahead: Cautious Optimism Amid Emerging Possibilities

8.5. Conclusions

9. References

Abbreviations

aCGH | Array Comparative Genomic Hybridisation | -

AI | Artificial Intelligence | Machine programmes and algorithms mimicking human cognition

ANOVA | Analysis of Variance | -

ART | Assisted Reproductive Technology

AUC | Area Under the (ROC) Curve | Metric quantifying overall model performance

BC | Blastocoel | Fluid-filled cavity within a blastocyst

BG | Background | Non-relevant portion of an image

b-hCG | Beta Human Chorionic Gonadotropin | Hormone indicating early pregnancy

CI | Confidence Interval

CNN | Convolutional Neural Network | Neural networks for grid-like data (e.g., images)

COC | Cumulus-Oocyte Complex

DNN | Deep Neural Network | Multi-layered networks for complex data representation

DSC | Dice Similarity Coefficient | Metric measuring overlap between segmentations

FHR | Foetal Heart Rate | Heartbeat of the developing foetus

GT | Ground Truth | Known correct labels for training/testing models

HEPES | 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

ICM | Inner Cell Mass | Inner cell cluster of a blastocyst

ICSI | Intracytoplasmic Sperm Injection

IVF | In Vitro Fertilisation

LH | Luteinising Hormone

LIN | Linearity

ML | Machine Learning | Algorithms learning from data without explicit programming

NGS | Next Generation Sequencing

OPU | Ovum Pick-Up

OR | Odds Ratios

PGT | Preimplantation Genetic Testing | Genetic testing of embryos before implantation

PGT-A | Preimplantation Genetic Testing for Aneuploidies | Screening embryos for chromosome number abnormalities

PGT-M | Preimplantation Genetic Testing for Monogenic Diseases | Testing embryos for single-gene disorders

PGT-P | Preimplantation Genetic Testing for Polygenic Disorders

PGT-SR | Preimplantation Genetic Testing for Structural Rearrangements | Detecting balanced chromosomal rearrangements in embryos

PN | Pronucleus

ROC | Receiver Operating Characteristic | Graph showing classification model performance

SA | Spontaneous Abortion | Miscarriage; loss of pregnancy before 20 weeks

SVM | Support Vector Machine | Classification algorithm finding optimal decision boundaries

TE | Trophectoderm | Outer cell layer of a blastocyst

TLM | Time Lapse Microscopy Imaging

VAP | Average Path Velocity

VCL | Curvilinear Velocity

VSL | Straight Line Velocity

ZP | Zona Pellucida | Glycoprotein layer surrounding an oocyte/embryo

Summary Table and Term Definition

Artificial Intelligence (AI): Computer systems/algorithms that can perform tasks requiring human-like intelligence, such as learning, problem-solving, visual perception, decision-making.

Machine Learning (ML): A subset of AI where algorithms learn and improve from data without being explicitly programmed. The computer determines optimal parameters to complete a task.

Deep Learning: A subfield of ML using artificial neural networks with multiple layers to automatically learn hierarchical features and representations from data.

Convolutional Neural Networks (CNNs): A type of deep neural network commonly used in image analysis for tasks like classification, object detection, and segmentation.

Computer Vision: Techniques and algorithms that enable computers to gain high-level understanding and extract meaningful information from digital images or videos.

Feature Extraction: The process of identifying and extracting relevant patterns or characteristics from images that are useful for subsequent analysis tasks.

Image Segmentation: Partitioning a digital image into multiple distinct regions or objects of interest, often to locate and identify relevant structures.

Image Representation: Encoding and representing images in a suitable mathematical format, such as vectors or matrices, to enable computational processing and analysis.

Ground Truth: The known outcomes or correct labels used to train and evaluate the performance of AI/ML models, serving as an objective reference.

Sensitivity: How well a model identifies true positives - the proportion of actual positive cases that are correctly predicted as positive by the model.

Specificity: How well a model identifies true negatives - the proportion of actual negative cases that are correctly predicted as negative by the model.

Overfitting: When a model learns the training data too well and loses the ability to generalise to new, unseen data. It memorises noise rather than learning general patterns.

Abstract

Defined as computer algorithms/software that can perform tasks that typically require human intelligence (e.g. learning, problem-solving, visual perception), Artificial Intelligence (AI) is rapidly growing, with vast applications across various fields. Regarding medical imaging, X-rays, MRIs, predictive modelling from patient data, ophthalmology, dermatology, radiology, pathology and assisted reproduction technologies (ART) all could benefit. In ART, the practice of IVF has the potential to benefit in several areas, including patient management, gamete/embryo identification, ranking and selection. This thesis aims to facilitate selecting and ranking IVF embryos for transfer using static images and AI analysis. In particular:

- To test prototypes for predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. This was achieved using an approach named AIR-e. The hypothesis that this novel approach is effective was accepted, indicating the feasibility of using AIR-e® to predict implantation potential from a single digital image.

- To test an Embryo Ranking Intelligent Classification Algorithm (ERICA), an AI clinical assistant with embryo ploidy and implantation predicting capabilities. This was achieved: the hypothesis that ERICA can be used to predict ploidy status in IVF embryos was accepted. Following training and validation, ERICA was more successful than random selection and experienced embryologists in ranking embryos with the highest implantation potential based on a static picture as the only source of information.

- To use ERICA to test the hypothesis that it can predict first-trimester pregnancy loss. In a retrospective pilot study, the hypothesis was accepted. Results support a correlation between the risk of spontaneous abortion and embryo rank as determined by ERICA.

- To develop the first automatic method for segmenting all morphological structures during the developmental stages of a human blastocyst. his was achieved. The approach can automatically segment blastocysts from different laboratory settings and developmental phases within a single pipeline. A sensitivity analysis established that this method is robust.

- To perform a systematic evaluation of predictive AI models in reproductive medicine. Here, a critical appraisal of AI models in reproductive medicine is discussed, conveying the importance of transparency and standardisation in reporting AI models so that the risk of bias and the potential clinical utility of AI can be assessed. Four reasons to interpret reproduction AI studies with caution are given, as is a guide to appraise AI's efficacy in reproductive medicine critically. Finally, a quick reference reader and referee consideration guide are provided.

AI's ability to make decisions based on facts and data makes the decision process reproducible and repeatable. AI can learn and analyse complex patterns at an increased resolution and with more variables far beyond most humans' capabilities. This thesis thus demonstrates that AI has the potential to be utilised as a promising tool to resolve many longstanding challenges in ART and assist clinicians in decision-making to achieve the ultimate goal of a healthy live birth. Similar principles may apply to sperm and egg analysis and other challenges in ARTs. Barriers include health record privacy terms, paper records and variations in electronic medical record systems. AI will play a significant role in the future of IVF that, realistically, will only be achieved if the artisanal approach of manual handling, basic microscopy and subjective analysis is replaced. The most likely combination of modalities to achieve scalability and improved access to IVF will be automation and AI.

1. General Introduction

1.1. Artificial intelligence and its impact on society

Artificial Intelligence (AI) is a rapidly growing field with numerous applications across industries. AI involves the development of computer systems in the form of algorithms and software that can perform tasks that typically require human intelligence, such as learning, reasoning, problem-solving, visual perception, speech recognition, language translation, decision-making, and perhaps even brainstorming (Ornes S, 2023). The applications of AI are vast, and they continue to expand as new technologies and developments emerge (Sahni NR, 2023)

1.1.1. AI in finance

An example of a significant application of AI is in the field of finance. AI algorithms can analyse vast amounts of data, including market trends, company performance, and consumer behaviour, to provide insights that help financial institutions make more informed investment decisions. AI can also be used to develop predictive models that can forecast market trends and identify potential investment opportunities (Dash, RK, 2023). Additionally, AI can be used to identify fraud and detect unusual financial transactions that may indicate criminal activity (Ryman-Tubb NF, 2018).

1.1.2. AI in industry

AI has been used in the transportation industry to improve safety and efficiency (Dong K, 2022). AI algorithms can analyse data from sensors and other sources to identify potential hazards and improve route optimisation. This can help reduce accidents and improve traffic flow. In addition, AI-powered autonomous vehicles that can drive themselves are being developed, potentially improving safety (Marti E, 2019).

In the manufacturing industry, AI has been used to improve production efficiency and quality control. AI algorithms can resource management tasks for systems and networking. This can help manufacturers reduce waste, forecast energy consumption, improve product quality, and increase productivity (Khalil M, 2022). In addition, AI-powered predictive maintenance systems can help manufacturers identify potential equipment failures before they occur, reducing downtime and maintenance costs (Barbado A, 2022, Rahman NHA, 2023).

1.1.3. AI in image analysis and processing

The field of image analysis and processing has witnessed significant advancements with the integration of AI techniques, enabling unprecedented accuracy and efficiency in image analysis and processing tasks, leading to improved accuracy and efficiency in image interpretation (Rajpurkar P, 2023). AI techniques employed for image analysis include deep learning, computer vision, and neural networks. Key components of AI-based image analysis include object detection, image recognition (Fig. 1.1), image segmentation, image enhancement, and image classification (Van der Velden, 2022). These components could also claim to be applications in their own right (Chen L, 2023).

FIGURE 1.1. Image of an intracytoplasmic sperm injection (ICSI), using computer vision and AI, displaying Object Detection: Tip of the pipette (green dot, extreme left) and the tip of ICSI needle (green dot, extreme right); and Image Recognition: oocyte (red square) and sperm (green square, extreme right). Image courtesy of Conceivable Life Sciences.

AI techniques, including deep learning, computer vision, and neural networks, form the foundation for image analysis and processing. Deep learning, a subset of machine learning, has gained prominence due to its ability to learn hierarchical representations from data automatically. Neural networks, particularly convolutional neural networks (CNNs), have become the cornerstone of image analysis tasks, demonstrating remarkable performance in object detection, image recognition, and other related tasks (Guo Y, 2016). Additionally, computer vision techniques, which focus on extracting meaningful information from images, complement AI algorithms by providing tools for image segmentation, feature extraction, and image understanding (Esteva A, 2021).

1.1.3.1. Deep learning

Deep learning is a subfield of machine learning that has gained significant attention and success in recent years, particularly in the field of computer vision (Chai J et al, 2021). It is a powerful approach that mimics the human brain's neural networks to automatically learn and extract meaningful representations from large amounts of data. The concept behind deep learning lies in the utilisation of artificial neural networks with multiple layers, known as deep neural networks, to learn hierarchical features from input data, such as images (Burkov A, 2019, Lancashire et al., 2009)

In the context of image analysis, deep learning has demonstrated remarkable capabilities in various tasks, including image classification, object detection, and image segmentation (Nixon M, 2019; Guo Y, 2016). By using convolutional neural networks (CNNs), a specific type of deep neural network, deep learning algorithms can automatically extract relevant features from images, enabling accurate and efficient analysis (Burkov A, 2019). This makes deep learning particularly useful in biology, as it can assist in tasks such as analysing microscopic images, identifying cell structures, and detecting patterns or anomalies (Kragh et al., 2019; Chavez-Badiola et al., 2020b). The power of deep learning lies in its ability to automatically learn and adapt to complex patterns and variations in image data, offering tremendous potential for advancing research and discovery in the field of biology.

1.1.3.2. Computer vision techniques

Computer vision techniques are critical to image analysis and processing, enabling us to extract meaningful information from images and understand their contents. Computer vision techniques can be immensely valuable for analysing microscopic images, identifying biological structures, and quantifying cellular phenomena (Gandomkar Z, 2018). These techniques involve a range of algorithms and methods that allow us to manipulate and interpret image data, including feature extraction, image representation, image filtering and image understanding (Rebouas Filho P, 2017; Tang Q, 2017; Guo X, 2014). Integrating with AI algorithms enables more sophisticated and accurate image analysis. (Cao P. et al, 2022; Guo Y, 2016)

1.1.3.2.1. Feature extraction

One fundamental aspect of computer vision is feature extraction (Fig. 1.2), which involves identifying and extracting relevant visual patterns or characteristics from images. This step is crucial for subsequent analysis tasks such as object recognition or segmentation. Feature extraction techniques can include edge detection, corners, texture analysis, and local feature descriptors, among others.

These methods enable us to capture and represent specific visual attributes, enabling more accurate and detailed analysis (Rebouas Filho P, 2017; Tang Q, 2017). Examples of feature extraction methods include SIFT (Scale-Invariant Feature Transform) and HOG (Histogram of Oriented Gradients) (Wamidh K, 2020; Nixon M, 2019); these aim to capture discriminative information that distinguishes different objects or structures in an image.

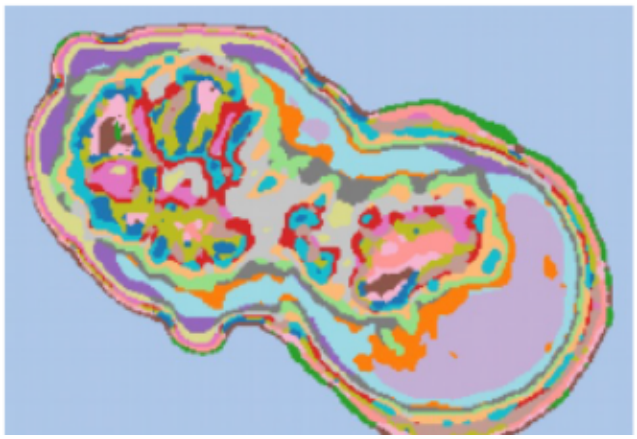


FIGURE 1.2. Feature extraction from a blastocyst stage embryo using an artificial vision filter.

1.1.3.2.2. Image segmentation

Image segmentation involves partitioning an image into distinct regions or objects of interest (Fig. 1.3). This process is essential for identifying and separating individual cells or structures within an image. Image segmentation algorithms can utilise various approaches, including thresholding, region-based methods, or advanced techniques like active contours or deep learning-based methods. Accurate segmentation enables further quantitative analysis and provides valuable insights into biological processes and structures (Nixon M, 2019).

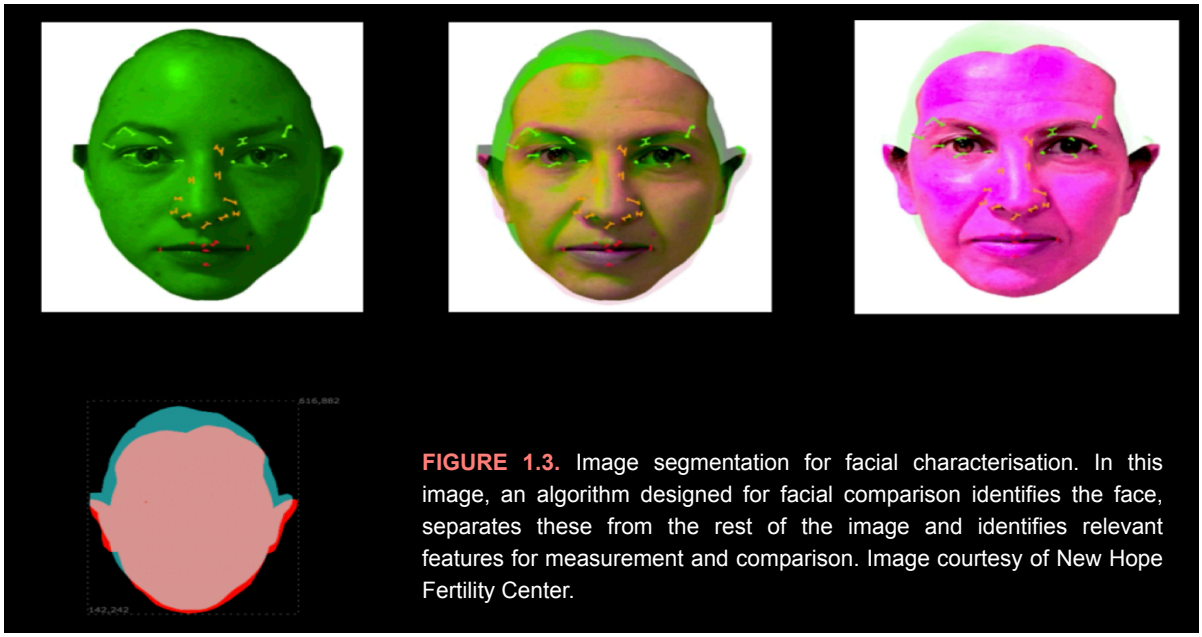


FIGURE 1.3. Image segmentation for facial characterisation. In this image, an algorithm designed for facial comparison identifies the face, separates these from the rest of the image and identifies relevant features for measurement and comparison. Image courtesy of New Hope Fertility Center.

1.1.3.2.3. Image representation

Image representation is the process of encoding and representing images in a suitable format for analysis. This can involve transforming images into mathematical representations, such as histograms, vectors, or graphs. By representing images in a meaningful and compact manner, computational algorithms can efficiently process and analyse the data. Common image representation techniques include colour spaces (Fig. 1.4), such as RGB (Red, Green, Blue) or HSV (Hue Saturation Value), and image descriptors, such as Bag-of-Visual-Words or deep features extracted from convolutional neural networks (Wamidh K, 2020; Nixon M, 2019).

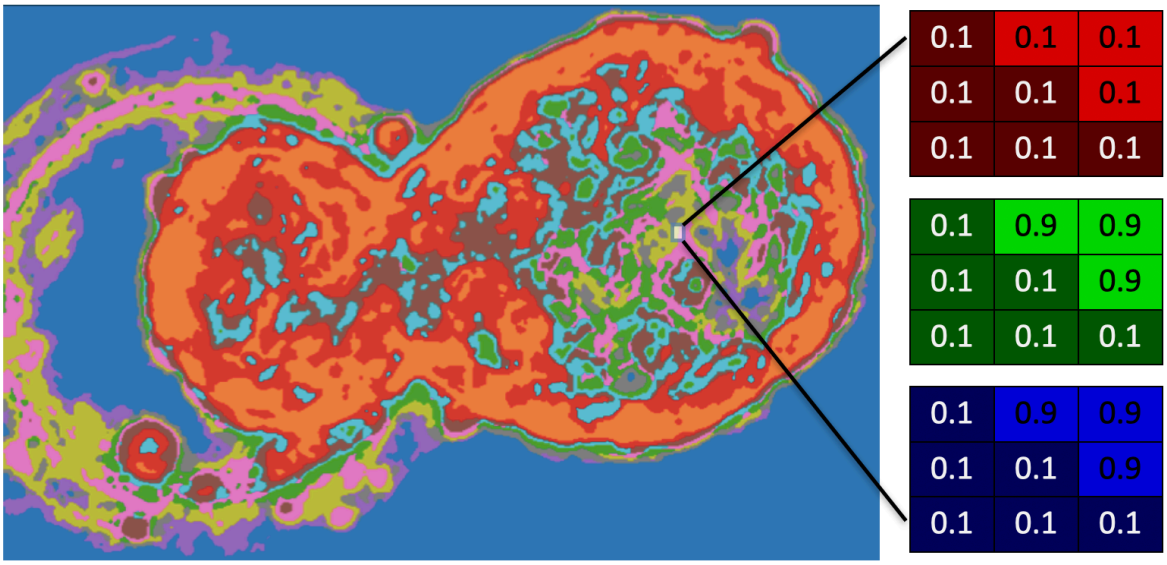


FIGURE 1.4. An example of RGB image representation from a pixel of a blastocyst image after applying an artificial vision filter.

1.1.3.2.4. Image filtering

Image filtering techniques are employed to enhance image quality, remove noise, or highlight specific image characteristics (Fig.1.5). Filtering operations can involve spatial or frequency domain operations, such as smoothing filters (e.g., Gaussian or median filters) or edge enhancement filters

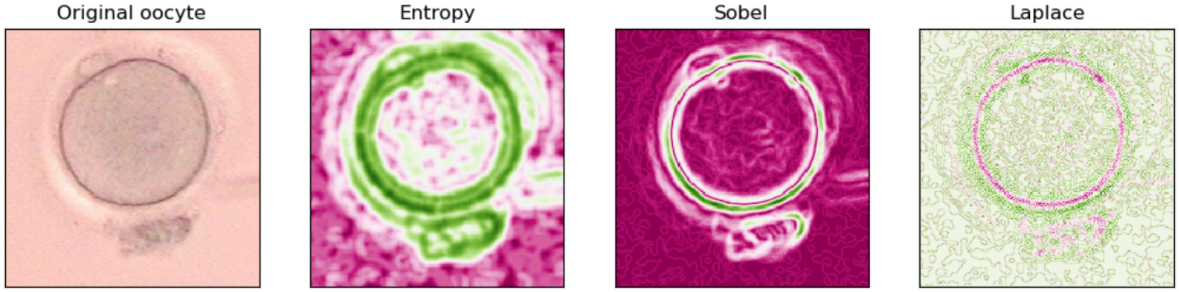


FIGURE 1.5. From left to right, an image of a human MII oocyte as seen under regular microscopy and through Entropy, Sobel and Laplace filters. Image courtesy of IVF 2.0 LTD.

(e.g., Laplacian or Sobel filters). These filters help to improve image clarity and enhance the extraction of relevant features for subsequent analysis (Nixon M, 2019).

1.1.3.2.5. Image understanding and image registration

Image understanding aims to extract higher-level semantic information from images, allowing for more advanced analysis and interpretation (Sarvamangala DR, 2022). This involves tasks such as object recognition, object detection, or scene understanding. AI algorithms, particularly deep learning approaches, have revolutionised image understanding by leveraging the power of neural networks to learn complex representations directly from the data (Farias AF, 2023). Convolutional neural networks (CNNs) have proven particularly effective in tasks such as image classification, object detection, and semantic segmentation (Guo Y, 2016).

Moreover, computer vision techniques can also involve image registration, where multiple images are aligned or fused, allowing for the comparison and analysis of different image modalities or time-lapse sequences. This is particularly useful in biology, as it facilitates studying dynamic processes or integrating data from multiple imaging techniques (Cocianu C et al, 2023).

Computer vision techniques form the bedrock of image analysis and processing. These techniques focus on extracting meaningful information from images, enabling comprehensive analysis and interpretation. We can significantly enhance image analysis and processing capabilities by integrating these computer vision techniques with AI algorithms (Sarvamangala DR, 2022, Mendizabal-Ruiz G, 2022). AI algorithms, such as deep learning models, can learn and extract features automatically from images, reducing the need for handcrafted features. This improves accuracy and allows for more robust analysis across different image domains and applications. Combining computer vision techniques with AI algorithms enables the development of sophisticated systems capable of performing tasks like automated image recognition, intelligent image retrieval, and advanced image-based decision-making in biology (Guo Y, 2016). These techniques enable us to uncover intricate details and patterns within images, contributing to a deeper understanding of biological systems and advancing research in many biological disciplines.

1.1.4. AI in medicine

The advances in AI have significantly impacted the healthcare industry. Using AI, vast amounts of medical data can be analysed to provide accurate diagnoses and treatment recommendations. AI

systems have been developed to analyse medical images, such as X-rays and MRIs, to identify and classify abnormalities accurately. This early detection of diseases can help doctors make more accurate diagnoses and improve treatment outcomes based on early intervention, which can significantly improve patient outcomes (Rajpurkar P, 2023). Specifically, and within the last ten years or so, machine learning, more specifically convolutional neural networks (CNN), has been used to assist with medical imaging in a variety of fields, such as ophthalmology (Abràmoff et al., 2016), dermatology (Esteva et al., 2017), radiology (Hosny et al., 2018) and pathology (Khosravi et al., 2018).

FIGURE 1.6. The use of AI in medicine makes headlines. From BBC News webpage (January 2020)

Another area where AI is transforming healthcare is in the analysis of patient data to develop predictive models. By analysing vast amounts of patient data, AI algorithms can predict disease progression and identify patients at high risk of developing certain conditions. Healthcare professionals can also use this information to develop targeted interventions, prevent disease progression and help improve outcomes (Chavez-Badiola, 2020a). Importantly, it could allow

clinicians to improve care by helping improve patient-doctor interactions (Brownstein JS, 2023; Haug CJ, 2023).

AI is also helping to improve the speed and accuracy of medical research by automating time-consuming tasks such as data collection and analysis. AI-powered tools can assist researchers in identifying potential new drugs, analysing clinical trial data, and identifying new disease biomarkers. This streamlined approach to medical research can speed up the development of new treatments and cures, benefitting patients and the healthcare industry. Today, it is difficult to find an area of medicine where AI is not being tested or developed to reduce costs, improve access and patient outcomes, and improve the job satisfaction of healthcare personnel (Haug CJ, 2023; Hickman C, 2020).

1.1.4.1. Introduction of AI models. Its effect on developers and users

Overall, the applications of AI are vast and continue to expand. As technology continues to evolve, AI will likely play an increasingly significant role in shaping the future of society. However, there are also concerns about the potential impact of AI on jobs and the economy. As AI technology continues to advance, it will be essential to consider the ethical and social implications of its use and ensure that it is used in a responsible and beneficial way (Haug CJ, 2023). Equally, every industry needs continuous improvement and optimisation of its processes to maintain and assure future competitive advantages in terms of better products and services, and cost-effective mechanisms to make the products and deliver the services to their clients. For this reason, it is expected that industries pursue to benefit from the most advanced technological tools that scientists and researchers are developing.

While most AI algorithms and computational methods may be challenging to implement, many open-source and free software libraries now ease the development process (Allen B, 2019; Widder DG 2022). However, the inclusion of these models into any clinical setting is a harder problem since developers have to work within a multidisciplinary team for a better understanding of the advantages, costs, requirements and limitations of the development process and set a risk management plan to ensure the identification of potentially hazardous scenarios on behalf of patients' safety (Rajpurkar P, 2022).

Potential users of AI models in Medicine include clinicians, embryologists, nurses, and administrative personnel. In this regard, user experience to ensure the newly introduced tools contribute to a more

efficient workflow becomes crucial since the potential clinical advantages of a system should not come at the cost of an increased workload (Hendrix N, et al., 2022).

Assuming that an AI has been successfully developed and its clinical recommendations are as good as those from an experienced senior healthcare provider, any user can access this “expert opinion” at any time without human biases, leading to suboptimal performance. These models, then, will have the advantages of being reproducible, scalable, tireless, and potentially reaching any place in the world, ideally at a low cost. This concept could dramatically change current practices in reproductive health, clinic administration, and the decision-making process internally and in front of the patient (Chavez-Badiola, 2022a, Hajirasouliha I, 2020, Rosenwaks Z, 2020).

1.1.4.2. AI and data sources

Some of the most recent and successful AI methods are based on the machine-learning paradigm on which the computer determines the optimal parameters of a high-dimensional model (mostly non-linear) used to complete the required task. These optimal parameters are computed employing numerical optimisation algorithms that attempt to minimise a defined error function that depends on the model's performance to complete the desired tasks evaluated using the examples provided (Burkov, A. 2019).

Therefore, the size and quality of the data used during the AI model's training are essential for its success. This is especially important in industries like reproductive medicine, where data is not as easily obtained due to data privacy and lack of structured EMRs (Electronic Medical Records). To overcome this problem, multi-centric collaborations are one way in which some groups have managed to increase the size of their data with excellent results (Hickman C., 2020). Low-quality and/or small-sized databases could result in biased models which might not be generalisable across clinics nor reproducible (Curchoe CL., et al, 2020a; Curchoe CL., et al, 2020b).

An AI model’s performance will be closely related to the database quality it was trained on. Although it might sound obvious, this is important when a user tries to interpret the results presented by an AI (Curchoe CL. et al., 2020a, Chavez-Badiola A, 2020c). This introduction will deal with the use of AI in assisted reproduction technology much later. Still, as the whole thesis is on this topic, an example is given: Let us say that the AI model “A” was trained with embryo images taken on day 3 using a good quality, multi-centric database to predict its probability of reaching blastocyst. However, a closer look

into the data might show that the entire database included embryos inseminated with intracytoplasmic sperm injection (ICSI) and grown in a sequential culture media. Is this model generalisable to images taken on day 3.5 or 4? Or is it accurate for IVF-inseminated oocytes? is it also useful for other types of culture media?

1.1.4.3. AI validation

In the case of machine learning algorithms, the validation of the technology consists of evaluating the model's performance with respect to a defined ground truth (Chavez-Badiola A, 2020a). Most of the time, this ground truth is defined by experts in the problem to be tackled, and the validation is carried out retrospectively. However, in some cases, validating the models in prospective studies is possible. In any case, the most common comparison of ML (Machine Learning) algorithms is the similarity in the performance with respect to what can be expected from one or more human experts (Curchoe CL, 2020a).

Also, since reproductive treatment can have many variations and paths, sensitivity analysis and general ability tests should be performed prior to the introduction of the technology in daily practice. Furthermore, before its clinical application, quality assurance should always be considered whenever an AI model is introduced into a new setting for the first time (Curchoe CL, 2020a; Chavez-Badiola A, 2020c).

1.1.4.4. Critical appraisal of an AI model

Reading about AI could be complex since it routinely introduces jargon alien to medical sciences and relies on sophisticated statistical and mathematical analyses, but still, most of the time, it’s presented within a clinical setting we are familiar with. However, there are specific guidelines which could help us to appraise AI technology better (Collins G, 2021; Liu Y, 2019), and efforts are being made to present recommendations for specialised practices, such as in Reproductive Medicine (Curchoe CL, 2020a).

AI users shall be trained to interpret an AI system's results and be aware of the limitations regarding the validation procedures and conditions of the model. To overcome this problem, a proper sensitivity analysis should be performed before depositing our trust in the AI model. However, suppose the user is able either to access a robust AI system to assure the quality of the results or to perform under similar conditions to those used for the training set. In that case, the user should perform a quality

assurance test with real data from their clinic before using the model in actual cases (Curchoe CL, 2020a).

As much as current medical practice is evidence-based oriented, we might have to acknowledge that the current concept of randomised controlled trials to prove efficacy and validation of an approach may not apply to AI and ML in the traditional sense. However, AI/ML are technologies with the potential to improve the results of fertility treatments by reducing the time to pregnancy and costs (Chavez-Badiola A, 2022b). Because of this potential, careful introduction of these technologies into clinical practice must be made through reflected decisions based on a basic understanding of the technologies at hand and the best ways to appraise them (Curchoe CL 2020a). To the best of my knowledge, there are no prospective studies published yet that assess the performance of an AI model compared to standard clinical decisions.

1.1.4.5. Ground truth and validation sets

As mentioned above, the “ground truth” consists of the known outcomes with which the AI is trained. Common outcomes in medicine are successful treatments (in reproductive medicine, this might be a positive pregnancy test, presence of heartbeat or born alive/healthy), known measurable status of the cells involved (e.g. genetic status, which in IVF might be ploidy status of an embryo or gamete), or any middle step indicating that the probability of a successful treatment is increased (e.g. in IVF this might be freeze-thaw survival of embryos, normal fertilisation, cleavage, blastocyst quality). AI systems are regularly trained to predict one of these outcomes that should be assessed in terms of usefulness in a “real-life” treatment protocol (Chavez-Badiola A 2022a). Also, the model shall be assessed using an independent validation set.

It is crucial to consider the database conditions (inclusion and exclusion criteria) and training size (after exclusions criteria) and assess if the whole spectrum of possibilities is included in the training and test set. This could guide a user in the product's restrictions or limitations. It is also possible that AI models were constructed over unbalanced data (true to false rate different than 1), which is also undesirable and could be a source of bias in the training process (Chavez-Badiola A 2020c).

Finally, it is always important to perform a “common sense check” on any results performed by AI. Perhaps take a few examples and see if they do indeed answer the questions asked. In the world of IVF, this may include morphological features of embryos that correlate to outcome or a measurable

feature such as implantation or ploidy status (Chavez-Badiola A 2020). The following sections deal specifically with reproductive medicine and the application of AI in the world of IVF.

1.2. Infertility, Assisted Reproduction Technology (ART) and In-Vitro Fertilisation (IVF)

Infertility is a multifaceted and emotionally challenging condition that impacts countless individuals and couples globally. It is characterised by the inability to achieve pregnancy after a year of regular, unprotected sexual intercourse (Practice Committee of the American Society for Reproductive Medicine, 2020). Infertility often carries a profound emotional toll for those experiencing it. The inability to conceive can evoke feelings of shame, guilt, and depression while also straining relationships, fostering social isolation, and diminishing overall quality of life (Ombelet W, 2011; Zegers-Hochschild F, 2009).

The prevalence of infertility varies across different regions, but the World Health Organization (WHO) estimates that approximately 10% to 15% of couples worldwide are affected by infertility. This equates to over 80 million couples facing infertility challenges. Notably, infertility can impact both men and women and can arise from various factors such as genetics, age, lifestyle choices, environmental influences, and medical conditions. The causes of infertility are often intricate and multifaceted, potentially involving one or both partners (Zegers-Hochschild F, 2009).

1.2.1 Causes of infertility

Infertility can be classified into four main categories: male factor, female factor, a combination of both or unexplained (idiopathic). Male factor infertility accounts for approximately 30% of cases, while female factor infertility is solely responsible for around 30-40% of cases. In the remaining instances, both partners may have contributing factors, or the cause of infertility remains unidentified (Carson SA 2021).

Male factor infertility can arise from various issues, including low sperm count, impaired sperm motility, and abnormalities in sperm morphology (shape). These factors can significantly impact the chances of successful conception (Okonofua FE, 2022; Carson SA, 2021).

On the other hand, female factor infertility can be attributed to a range of causes. Ovulatory disorders, such as irregular or absent ovulation, can hinder the release of mature eggs necessary for fertilisation. Tubal and pelvic factors, such as blocked or damaged fallopian tubes, can obstruct the passage of eggs and sperm. Uterine factors, such as abnormalities in the structure of the uterus or the presence of fibroids, can also interfere with implantation and pregnancy. It is important to note that infertility can result from a combination of male and female factors, making it a complex and multifaceted issue. In some cases, despite thorough investigations, the precise cause of infertility may remain unknown (Carson SA 2021).

Advanced age plays a crucial role in infertility, particularly among women. Women are born with a limited supply of eggs, and both the quantity and quality of eggs diminish over time. This decline in egg quality can result in reduced fertility, an elevated risk of miscarriage, and an increased likelihood of chromosomal abnormalities in their offspring (Wei L, 2023; American College of Obstetricians and Gynecologists, 2020). Additionally, lifestyle factors like smoking, excessive alcohol consumption, and obesity can further contribute to infertility (Bala R, 2021; Hazlina N, 2022). Gaining knowledge about the causes of infertility and being aware of available treatment options is crucial for individuals and couples facing difficulties in conceiving.

1.2.2 Treatments for infertility

The treatment of infertility depends on the underlying cause of the problem and may involve medications, surgical procedures, or different forms of assisted reproduction technologies (ART). These treatments have revolutionised the management of infertility, with IVF profiling itself becoming the most commonly used assisted reproductive technology (Doody K, 2021).

Fertility drugs, along with lifestyle changes, are often the first line of treatment for women who have ovulation issues. These drugs work by stimulating the ovaries to release eggs and can increase the chance of conception. Success rates for fertility drugs vary depending on the specific medication used and the underlying cause of infertility, but overall success rates per cycle remain low (Carson SA 2021).

In cases where fertility drugs are not effective, ART may be used. ART includes a variety of procedures, such as intrauterine insemination (IUI) and IVF. IUI involves placing sperm directly into

the uterus, while IVF involves fertilising eggs outside of the body and then transferring the resulting embryos into the uterus (HFEA, 2023).

The success rates of infertility treatments vary depending on several factors, including age, the underlying cause of infertility, and the type of treatment used. For instance, the success rate of IUI ranges from 5% to 15% per cycle (Bahadur, G, 2020). In comparison, IVF has a higher success rate, with live birth rates per cycle ranging from 40-50% for women under 35 years of age, reducing to 10-15% per treatment cycle for women over 40 (Bahadur G, 2020; HFEA 2021; SART 2021).

1.2.3. In-vitro fertilisation (IVF) and intra-cytoplasmic sperm injection (ICSI)

In the early 20th century, scientists began experimenting with fertilising eggs outside the body. However, it wasn't until the 1970s that Robert Edwards and Patrick Steptoe achieved the first successful IVF (Steptoe P, 1978). The technique involved retrieving eggs from a woman's ovaries, fertilising them with sperm in a laboratory dish, and then transferring the resulting embryo(s) to the uterus. However, IVF success rates were initially low, prompting researchers to develop new techniques to improve fertilisation rates (Niederberger, 2018). One such technique was intracytoplasmic sperm injection (ICSI), which was first successfully used in humans in 1992 by Gianpiero Palermo in Belgium (Palermo G, 1992). This technique involves injecting a single sperm directly into the cytoplasm of the egg, bypassing the need for the sperm to penetrate the outer layer of the egg on its own. Currently, IVF and ICSI are widely used fertility treatments, with ongoing research and development aimed at improving their success rates and minimising potential risks. Both are highly complex and sophisticated assisted reproductive technologies that have revolutionised the field of reproductive medicine. IVF (with or without ICSI) is a multi-step process that involves several key stages, each carefully managed and monitored to optimise the chances of a successful outcome (Zhang, 2016).

The first step in the IVF/ICSI process is ovarian stimulation (Fig 1.7). Ovarian stimulation entails administering medication over two weeks to induce the production of multiple mature eggs, with evaluations conducted regularly to monitor ovarian response and identify the optimal time for egg retrieval. These evaluations are crucial in fine-tuning the medication dosage to maximise the yield of high-quality oocytes while minimising the risk of complications. This stage aims to obtain as many top-quality eggs as possible, increasing the likelihood of fertilisation and the development of healthy embryos (Zhang, 2016).

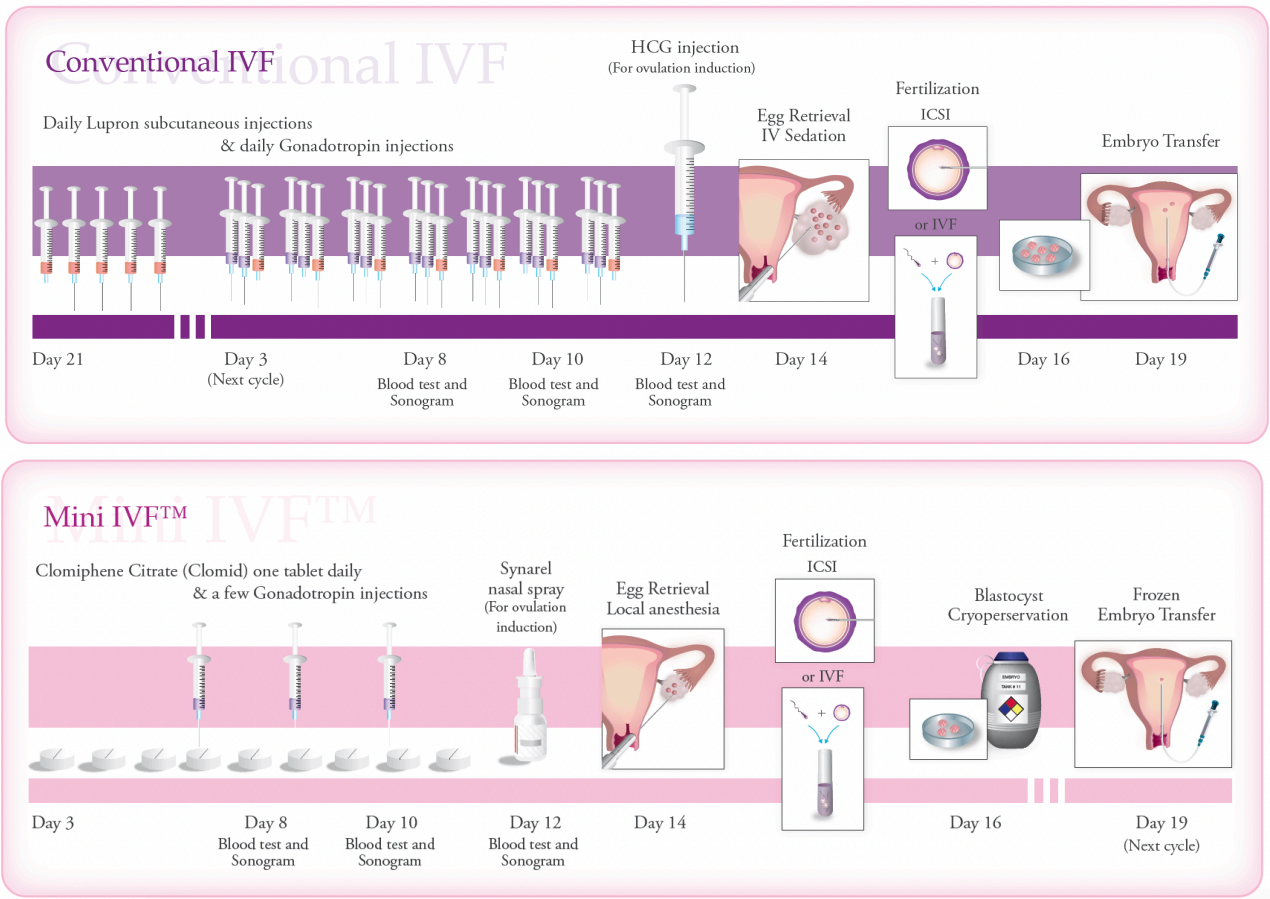


FIGURE 1.7. The IVF process, from ovarian stimulation to embryo transfer using two different stimulation protocols: Conventional and Minimal Stimulation. The latter includes blastocyst cryopreservation and the need for a second cycle dedicated to embryo transfer. Images courtesy of Daniela Leroy Murillo and New Hope Fertility Center.

Once the eggs have been retrieved from the ovaries, embryologists carefully examine them under a microscope to assess their quality and maturity. The embryologists check for a polar body and, in some cases, the meiotic spindle, indicating that the egg has completed its first meiotic division and is ready for fertilisation (Chavez-Badiola A, et al., 2017). They also evaluate the size and shape of the egg, as well as the thickness of its zona pellucida, which is a protective layer surrounding the egg. After egg evaluation, the mature eggs are then prepared for fertilisation. There are two main methods of fertilisation used in IVF: conventional insemination and intracytoplasmic sperm injection (Doody, 2021).

In conventional insemination, a concentrated sperm sample is added to each dish containing the eggs, and they are left to incubate overnight. The sperm then swim towards and penetrate the outer layer of the egg. In contrast, ICSI involves injecting a single sperm directly into the cytoplasm of the egg using a micro-needle (Palermo, 1992). This technique is used when the quality or quantity of sperm is poor, when previous fertilisation attempts using conventional insemination have failed, or

when special studies, such as pre-implantation preimplantation genetic testing (PGT), or techniques are required (Doody, 2021; Zhang, 2017; Munne, 1993).

In ICSI, the process of sperm selection is crucial to ensure the best possible outcome. For this, the embryologist selects the most viable-looking sperm, typically using high-magnification microscopy to assess the sperm's shape, motility, and other characteristics (Mendizabal-Ruiz G,2022). The selected sperm is then loaded into a fine glass needle and carefully inserted into the cytoplasm of the mature egg (Palermo, 1992).

After the fertilisation process, the embryos are closely monitored for development and quality by embryologists. The embryos are checked regularly to ensure they grow properly and have the correct structure and number of cells. Embryos are typically graded based on their appearance and developmental stage, and the best-quality embryos are then selected for transfer to the uterus or cryopreservation for later use (Zhang JJ, 2017; Kuwayama, 2007).

The last stage of the IVF/ICSI process is the transfer of one or more embryos to the uterus, which is done using a catheter through the cervix. The number of embryos transferred is determined by several factors, such as the woman's age and the quality of the embryos. It is ideal to transfer only one embryo to avoid multiple pregnancies, but in some instances, more than one embryo may be transferred (Gliozheni, 2023).

Following embryo transfer, progesterone is usually administered to the woman to support the luteal phase and to allow for the pregnancy's growth and development. A pregnancy test is usually conducted a week after whether the treatment has been successful (Zhang JJ, 2016).

IVF/ICSI is a highly effective treatment for infertility, with variable success rates depending on a range of factors, including the age of the woman and the quality of the embryos. While it can be a time-consuming and emotionally challenging process, it offers many couples the best chance of achieving a successful pregnancy and starting a family (Doody, 2021).

Achieving a successful pregnancy with the minimum of treatment intervention is the aim of every infertile couple and fertility clinic. Unfortunately, factors affecting pregnancy are multifactorial and incompletely understood. In addition to female age (and subsequent egg quality), over 200

confounders affect outcomes in assisted reproduction (Pool, 2012), the majority stemming from the quality of gametes and embryos and the effects of the laboratory and procedures. No single test, observation, or algorithm has been found to predict implantation (Chavez-Badiola, 2020a). Abnormal embryo genetics, the most common cause of miscarriage (Popescu, 2018), as well as several other factors, including staff effects, the culture environment, endometrial environment, maternal and paternal health, ease of embryo transfer, will always have some influence on the outcome (Sciorio, 2023); therefore, it is improbable that any embryo assessment can currently guarantee 100% prediction of a successful outcome.

1.2.4. Morphological analysis of IVF/ICSI embryos

1.2.4.1. Classic analysis

Since the inception of IVF, embryo morphology has been the gold standard for embryo selection, and this method is still the most accepted and commonly used to date. Despite clear and detailed guidelines for morphology-based classifications, embryo morphology assessment remains an artisanal technique lacking objectivity and reproducibility, limiting its overall predictive value (Chavez-Badiola 2020b; Adolfsson and Andershed, 2018). Furthermore, current scoring systems for human blastocysts focus on only 3 parameters, i.e., degree of expansion, quality of inner cell mass and quality of trophectoderm (Gartner 2000); other features that may have an impact on development potential are essentially disregarded, limiting its predictive potential (Yoshida 2018; Sciorio 2018).

1.2.4.2 Time-lapse and morphokinetics

Recent advances have enabled the acquisition of embryo images at regular time intervals without disrupting culture conditions, thanks to the introduction of time-lapse incubators (Kovacs 2014). This allows for observation of individual embryos' developmental stages as they occur, providing an abundance of data points to inform embryo selection (Hinojosa 2022). However, the impact of this approach on improving IVF outcomes remains uncertain (Armstrong et al., 2019). Time-lapse incubators with integrated imaging systems come at a high cost, and results analysis may be time-consuming, depending on the assessment depth (Chen 2017, Conaghan 2013, Kaser 2016, Kirkegaard 2015, Sciorio 2021).

The application of time-lapse monitoring in IVF has been a topic of debate in recent years. Several studies have suggested that time-lapse monitoring, coupled with the use of morphokinetic algorithms

for embryo selection, can lead to improved clinical outcomes compared to conventional single time-point morphology assessment. A meta-analysis by Pribenszky et al. (2017) found that time-lapse monitoring was associated with significantly higher ongoing clinical pregnancy rates, lower early pregnancy loss, and increased live birth rates. However, the quality of evidence in this meta-analysis was moderate to low due to inconsistencies across the included studies and the selective application of the technology. Additionally, concerns have been raised regarding the study inclusion/exclusion criteria and potential conflicts of interest (Armstrong et al., 2018; Alikani, 2018).

Further studies have investigated the use of time-lapse imaging morphokinetic-based algorithms for ranking blastocysts and predicting live birth outcomes. Fishel et al. (2018, 2019) found that these algorithms provided objective hierarchical ranking of embryos and demonstrated superior discriminating power compared to conventional blastocyst morphology assessment. However, it is important to acknowledge that these studies were retrospective and prospective in nature, and further well-designed studies are needed to establish the efficacy and clinical utility of time-lapse monitoring (Basile, 2019). Despite the controversies surrounding time-lapse technology, its potential benefits in improving embryo selection and overall success rates in IVF cannot be ignored, and ongoing research and development in this field, including the use of automation and artificial intelligence, may help to address current limitations and optimise its application in clinical practice.

Automated annotation of key events in embryo development and incorporating artificial intelligence algorithms appear to be sensible approaches for enhancing the capabilities of time-lapse embryo evaluation (VerMilyea 2020, Malmsten 2020). This could allow for identifying new and possibly more pertinent evaluation metrics, ultimately improving how predictive the results are compared to human assessment (Bamford 2022, Fishel 2020).

1.2.5. Preimplantation genetic testing (PGT)

1.2.5.1. PGT for the diagnosis of genetic disease

The earliest reported application of PGT was to allow for genetic testing to enable sex selection for a couple at risk of transmitting a monogenic disease, in more detail, an X-linked single gene disorder (Handyside 1990, Handyside 1989). Initially used for sexing, then cystic fibrosis in 1992 (Handyside 1992), progress was traditionally impeded by the need to tailor each diagnosis to the mutation. Universal approaches such as Karyomapping (Handyside 2010) have become the gold standard for

preimplantation testing for monogenic diseases (PGT-M). Karyomapping allows the detection of the inheritance of (grand) parental haplotype blocks through the tracking of inherited chromosomal segments. It involves genome-wide single nucleotide polymorphism (SNP) analysis of parental DNA, a reference-related individual of known disease status (typically an affected child) and amplified DNA from embryo biopsy. The Karyomapping livery displays homologous chromosomes, points of crossing over and the haplotype of each of the embryos. Karyomapping also detects chromosome disorders for preimplantation testing for aneuploidies (PGT-A) (Natesan 2014).

In PGT-M, targeted diagnoses are an intrinsic part of the analysis (Daar 2018). In contrast, preimplantation testing for polygenic diseases (PGT-P) is more complex, as it involves testing multiple genes that contribute to the condition for with a polygenic risk score (PRS) (Treff 2020, Treff 2019, Sudlow 2015) established by a genome-wide association study. PGT-P has benefits and drawbacks, with vocal proponents and opponents on both sides (Neuhausser 2023, Forzano 2022). The question often raised is how much benefit PGT-P gives patients compared to the inherent risks. For each disease, it has been suggested that a case would have to be made for PGT-P, as will a robust justification that the couple involved would, in fact, actually benefit (Forzano 2022).

Preimplantation genetic testing for structural rearrangements (PGT-SR) is the least well-known variant of PGT that provides effective treatment for many couples at risk of transmitting chromosome disorders by carrying a balanced translocation. Structural chromosomal rearrangements usually result in infertility, repeated implantation failure, pregnancy loss and/or affected children despite the parent carrier having no apparent phenotype (Ogur 2023). They nonetheless produce chromosomally unbalanced gametes and embryos – hence the need for PGT-SR to select for those that are karyotypically normal or balanced. PGT-SR largely uses array CGH, SNP arrays, Karyomapping and, most recently, next-generation sequencing (NGS) following trophectoderm biopsy (Tong, et al., 2022). The greatest limiting factor is the availability of normal or balanced embryo(s) for transfer, which can be affected by rearrangement type, the chromosomes involved, the sex of the carrier parent and *de-novo* aneuploidy, particularly if advanced maternal age is an issue (Griffin 2018). The efficacy of PGT, in all its variations, has been the subject of debate for over 20 years (Forzano 2022, Iews 2018).

1.2.5.2. PGT-A for the improvement of IVF success

PGT-A is, perhaps, the most widely indicated form of PGT. This approach was developed to improve IVF success as well as avoid miscarriage and live-born aneuploid offspring (Niederberger 2018). Diagnostic approaches include array comparative genomic hybridisation (aCGH), quantitative PCR, next-generation sequencing (NGS) and Karyomapping. Non-selection trial data provides evidence of PGT-A's efficacy in improving live birth rates per treatment cycle (Sanders 2021, Griffin 2017). In contrast, randomised controlled trial data is mixed, with no demonstrable benefit for cumulative pregnancy rates (Cornelisse 2020). Mosaic embryos (with a mixture of diploid and aneuploid cells) are a confounding factor, often arising post-zygotically and with evidence of live birth rates being reduced but not zero when transferred (Viotti 2023). PGT-A continues to attract heated debate, prompting questions of when one should consider the evidence base sufficiently robust to justify the routine use of PGT-A (Munné et al., 2019; Weissman et al., 2017).

Drawbacks to PGT-A are the invasive nature, cost, requirement to freeze all embryos, high follicular response, and an assumption of diagnostic accuracy of the few cells extracted from the embryo. Future developments include minimally invasive PGT-A, such as blastocoelic fluid assessment, non-invasive IVF, *i.e.* PGT-A from analysing spent culture media, and whole-genome assessment. At present, these techniques are still being researched (García-Pascual 2023).

1.2.6. The relationship between embryo morphology and aneuploidy

Embryo time-lapse systems are often employed in IVF centres with the aim of predicting embryo viability and improving the chances of unaffected live births. An ongoing question is whether aneuploid embryos display differing morphokinetic properties to their euploid counterparts. The rationale for PGT-A is that aneuploidy is the leading cause of IVF failure and miscarriage. Thus a non-invasive morphologic or even morphokinetic approach to stratify embryos for ploidy status would represent a great advance. This would avoid or minimise the costs of PGT-A. Factors such as fragmentation, multinucleation, abnormal cleavage and contraction have all been looked at. In a systematic review, Bamford et al. (2022) analysed 58 studies from >40,000 embryos, finding 10 morphogenetic variables significantly altered in aneuploid *vs* chromosomally normal embryos. These included time to eight cells, time to formation of a full blastocyst, and time to expanded blastocyst. They also noted some prognostic potential in the degree of fragmentation, multinucleation persisting to the four-cell stage and frequency of embryo contractions. No association with early multinucleation or unequal cleavage were found. Notably, widespread variability exists among aneuploid and euploid

embryos; thus, definitive classification is not currently feasible. Algorithms involving AI for live births may have some predictive ability (Bamford 2023, Bamford 2022).

1.3. The need for AI in ART (specific reference to image-based evaluation)

As mentioned in section 1.1, efforts directed at improving the accuracy and standardisation of image analysis through the development of computer-aided tools have recently gained attention in many medical fields (Gandomkar 2018, Rebouças Filho 2017, Tang 2017), including dermatology and oncology, in which linguistic data mining, deep learning techniques and architectures are used for image analysis and to assist diagnosis (Guo 2014, Hu 2018). There are many reasons why AI inclusion in reproductive medicine is highly desirable. As an example, AI's ability to make or to assist decisions based on facts and data that support them, makes the decision process reproducible and repeatable, in contrast to humans, whose decisions may depend on emotional states and are prone to subjectivity, fatigue, and other types of biases. AI can learn and analyse very complex patterns at an increased resolution or higher number of variables that are far beyond the capabilities of most human beings (Chavez-Badiola 2020a).

The field of reproductive medicine and its ventures into the world of AI has, as yet, had a rather limited clinical impact (Goodman 2016). Evaluating embryo characteristics with digital imaging systems based on quantitative parameters has become an active research area (Lagalla 2015, Rocha 2017a). Efforts have been made to improve objectivity during the embryo selection process by combining multiple morphologic parameters and morphokinetic analysis in mathematical models to improve predictive value for development potential and implantation. Such models, however, are often limited to the initial development stages of the embryo, dependent on currently existing classifications, or require sophisticated time-lapse microscopy systems (Mio 2006, Storr 2015, Majumdar 2017).

AI, an emerging technology aimed at embryo selection, may have additional applications, such as determining the diagnostic accuracy of PGT (Liang 2019). In medical practice, AI is becoming more accepted and promises the ability to analyse vast data sets at speeds not possible for humans to compute in the accepted time frame of preimplantation embryo culture. The advanced mathematics lends itself well to a static image or video analysis of embryo development and can compute and

analyse cryptic morphological features. By nature, it is non-invasive and, in due course, may be affordable and more accurate than established approaches as more data is evaluated and compared between different sources. Embryo grading by conventional morphology evaluation and morphokinetics analysis may not be the best or most accurate way to predict implantation potential, but it is standardised and recognised globally (Bormann 2020).

1.3.1. AI and gametes

Both invasive and noninvasive methods are used to select competent, healthy gametes for combination during assisted reproductive technology (ART) procedures. Every stage of ART treatment (fertilisation, embryo development, implantation, healthy clinical pregnancy) depends on high-quality, mature, genetically normal sperm and oocytes. Morphology of oocytes (cumulus-oocyte complex, polar body, and ooplasm defects) and motility characteristics of sperm (swim up, gradient centrifugation or laminar flow micro-channels on a chip, and PVP challenge) combined with morphology (vacuoles, head shape, and midpiece and tail defects) are routinely used to select gametes for insemination. Unfortunately, developmentally incompetent oocytes may exhibit the same morphology as competent ones, and even high-powered microscopy like IMSI (Intracytoplasmic morphological sperm injections) cannot detect DNA fragmentation in sperm. AI is perfectly suited to analyse and detect these defects better than the human eye, even aided by microscopy. Another approach has explored using AI in combination with immunogenetics data (HLA) to predict recurrent miscarriage (Moran-Sanchez 2019). This research suggests that accurately assessing the risk of recurrent miscarriage associated with a given pair of gametes could improve gamete donor selection and, therefore, increase pregnancy success rates.

1.3.1.1. AI and reproductive urology

In reproductive urology, early AI applications focused on semen parameters, but the technology has advanced to include the development of automated sperm detection, semen analysis, and disease prediction methods. AI technology for semen analysis, sperm viability, and DNA integrity has been bridged with external hardware devices and smartphone (mobile) applications (Dimitriadis 2019a, Kanakasabapathy 2017).

Semen analysis is a time-consuming, subjective, and labour-intensive process. It takes embryologists many years to develop the judgment to select sperm and the competence to perform intracytoplasmic sperm injection (ICSI) that produces very high rates of chromosomally normal BLs. Additionally, many embryologist hours are needed to find viable sperm for ICSI in certain surgical sperm extraction (TESE/TESA) samples, as well as to confirm the absence of sperm after a vasectomy. AI systems will likely perform these tedious, routine, and time-consuming tasks more accurately, if not much faster. Despite advances like computer-assisted semen analysis, automation of semen analysis has not been widely adopted, either because of cost or accuracy, and is still mostly performed by individual andrologists. AI technologies to predict which sperm is most likely to be chromosomally normal during ICSI or to identify a rare cell in a sample of mixed cell types have yet to be developed. However, commercial entities are now trying to solve these problems with promising results (Chavez-Badiola 2022).

Reviewing the available literature, I see that in 2014, Sahoo et al. used data mining techniques to predict seminal quality in men, taking into consideration lifestyle (smoking and alcohol consumption) and environment (childhood diseases and fevers) factors to achieve varying model accuracy (MLP: 92%, SVM: 91%, SVM+PSO 94%, NB: 89%, DT: 89%) to predict male fertility and to assess the importance of each feature employed to determine infertility (Sahoo 2014). Mirsky and colleagues employed interferometric phase microscopy along with support vector machine classifiers (SVM) to develop a model to assess sperm morphology and classify sperm into “good” or “bad” morphology with over 88% accuracy (Mirsky 2017). Kanakasabapathy and colleagues used smartphone microscopy with deep transfer learning to develop an inexpensive system that accurately measures sperm morphology based on Kruger's strict criteria (Kanakasabapathy 2017).

Akinsal et al. trained an ANN to analyse age, height, testicular volume, ejaculate volume, FSH, LH, and testosterone to predict the presence of chromosomal abnormalities in azoospermic men with >95% accuracy (Akinsal 2018). This method can identify the patient population most likely to need genetic testing.

1.3.1.2. AI and the impact of controlled ovarian stimulation on oocytes

Liao et al. have shown that an ML-derived algorithm is useful to assist clinicians in making an efficient and accurate initial judgment on the condition of patients with infertility. In their study, over 60,000 infertile couples’ medical records were evaluated using a grading system that classified the

patients into 5 grades ranging from A to E. The worst grade, E, represented a 0.90% pregnancy rate, while the pregnancy rate in the A grade was 53.8%. The cross-validation results showed the system's stability was 95.9% (Liao 2020).

Letterie et al. evaluated a computer decision support system for managing ovarian stimulation during IVF following key decisions made during an IVF cycle: [1] stop stimulation or continue stimulation. If the decision was to stop, the next automated decision was to [2] trigger or cancel. If the decision was to continue stimulation, then the following key decisions were [3] the number of days to follow-up and [4] whether any dose adjustment was needed (Letterie 2020). The authors used data derived from an electronic medical records system of a female population undergoing IVF cycles and oocyte cryopreservation to include the patients' demographics, past medical history, and infertility evaluation, including diagnosis, laboratory testing for ovarian reserve, and any radiologic studies pertinent to a diagnosis of infertility. The four key decisions during the ovarian stimulation and IVF process were compared to expert decisions across 12 providers; they were found to have a sensitivity of 0.98 for trigger and 0.78 for cycle cancellation.

Controlled ovarian stimulation (COS) yields oocytes at various stages of meiotic maturity. Identification of MII (extruded polar body), MI (no polar body), GV (germinal vesicle indicative of prophase I), giant MII oocytes, and other abnormalities is primarily performed by embryologists; however, nuclear and cytoplasmic maturity cannot be assessed. Noninvasive AI methods to evaluate oocyte competency could become an important selection and prediction tool to reduce the number of embryos created and wasted (of paramount importance in countries that restrict supernumerary embryos), to reduce the number of embryos for trophectoderm (TE) biopsy and PGT, and to prognose the success of an IVF cycle. In the case of donor egg cycles, a tool to objectively assess oocyte quality and subsequent fertilisation potential may be very valuable to intended parents for psycho-social reasons. Additionally, experimental and research procedures like in vitro maturation (IVM) of oocytes, somatic cell nuclear transfer and reprogramming, in vitro gametogenesis (IVG), and more would benefit from prediction and selection AI systems.

In 2011, Setti and colleagues performed a meta-analysis to identify the relationship between oocyte morphology and ICSI outcomes (Setti 2011). Their study demonstrated that a large first polar body and a large perivitelline space and the inclusion of refractile bodies or vacuoles are associated with decreased oocyte fertilisation. In 2013, Manna et al. performed a texture analysis of 269 oocyte

images and tracked the corresponding embryo development (Manna 2013). Texture features were used with a neural network to predict the outcome of a given cycle, meaning that multiple transfers were present in the data used for an AUC of 0.80. In 2021, Targosz and colleagues tested 71 deep neural network models for semantic oocyte segmentation (Targosz 2021). They trained their algorithm to classify the following oocyte morphologic features: clear cytoplasm, diffuse cytoplasmic granularity, smooth endoplasmic reticulum cluster, dark cytoplasm, vacuoles, first polar body, multi-polar body, fragmented polar body, perivitelline space, zona pellucida, cumulus cells, and the germinal vesicle. In this study, the top training accuracy (Acc) reached about 85% for training patterns and 79% for validation ones when using one of the variants of the DeepLab-v3-ResNet-18 model.

In 2020, Kanakasabapathy and colleagues trained a CNN to predict fertilisation (2PN or non-2PN) potential from oocyte images and to identify oocytes with the highest fertilisation potential >86% of the time (Kanakasabapathy M 2020). Results from this study allow for the development of novel quality assurance tools used to monitor oocyte stimulation regimens, assess ICSI performance, maintain optimal fertilisation and embryo culture conditions, and evaluate oocyte vitrification and warming procedures. That same year, Dickinson and colleagues used deep CNNs to locate the first extruded polar body and to identify the correct location on the oocyte to inject sperm for ICSI. In their study, over 14,000 images of MII oocytes were used for training, validation, and testing. The deep learning CNN was able to correctly identify the location of the polar body and the corresponding location for sperm injection for a test set of 3,888 oocytes with 98.9% accuracy with a 95% confidence interval (CI) (Dickinson 2020).

1.3.2. AI and embryo assessment

By far, the most interesting area of interest in AI as applied to IVF is the use of image analysis of IVF embryos to predict reproductive outcomes (Dimitriadis 2022).

1.3.2.1. Pronuclear-stage assessment

Normal fertilisation follows an apparently definite course of events. Oocytes show circular waves of granulation within the ooplasm after ICSI. During this granulation phase, the sperm head decondenses and the second polar body is extruded. The formation of the male pronucleus follows this. Simultaneously, the female pronucleus forms and is drawn toward the male pronucleus until

apposition is achieved. Both pronuclei then increase in size, and their nucleoli move around and arrange themselves near the common junction (Tosti 2016). Only zygotes with two distinct pronuclei are considered normal and appropriate for transfer. It is critical that embryologists assess fertilisation status correctly, as there is only a small window in which pronuclei can be properly counted.

Fertilisation checks and embryo quality assessments require manual examination, status recording, and embryo development scoring. These processes are labour-intensive and subjective. In 2019, Dimitriadis and colleagues described the development of a CNN that can distinguish between 2PN and non-2PN zygotes at 18 hours post-insemination with >90% accuracy (Dimitriadis 2019b). This system can be used as an embryologist's aid to help confirm the fertilisation assessment of each oocyte. It can also monitor individual embryologists performing ICSI in a clinical setting for advanced quality assurance to improve patient outcomes (Bormann 2021).

At the pronuclear stage, morphologic scoring, including alignment and size of pronuclei, alignment and number of nucleoli, and halo effect of the cytoplasm, can serve as a valuable noninvasive selection method in addition to the assessment of morphologic characteristics on the day of transfer. The morphologic features of the pronuclear zygote are related to implantation and pregnancy rates (Scott 1998, Scott 2000, Tesarik 1999). Pronuclear evaluation, therefore, is considered helpful in determining the most suitable embryos for transfer, thus achieving an optimal chance of conception (Lan 2003, Zollner 2003). Manually scoring zygotes is a labour-intensive and subjective activity. As such, few practices continue to assess this critical stage of development. However, using AI, these predictive features may be readily incorporated into an embryo selection algorithm (ESA). In 2021, Zhao and colleagues used CNNs for the segmentation of pronuclear-stage embryos. They examined the morphokinetic patterns of the zygote cytoplasm, zona pellucida, and pronuclei. Their manually annotated test set had a precision of >97% for the cytoplasm, 84% for the pronuclei, and approximately 80% for the zona pellucida. The authors concluded that their CNN system has the potential to be applied in practice for pronuclear-stage human embryo segmentation as a robust tool with high precision, reproducibility, and speed.

1.3.2.2. AI and cleavage stage-assessment

Embryos are at the cleavage stage 2-3 days after fertilisation and reach the blastocyst (BL) stage 5-7 days after fertilisation. Cleavage-stage embryos are generally selected for transfer based on cell number, degree of cellular fragmentation, and the overall symmetry of the blastomeres (Prados

2012). Traditional methods of embryo selection rely on visual embryo morphological assessment and are highly practice-dependent and subjective. Advances in time-lapse imaging (TLI) techniques have enabled regular and automated data acquisition of embryo development under controlled environments, along with identifying objective morphokinetic parameters (Azzarello 2012, Cruz 2012, Hlinka 2012, Lechniak 2008, Lemmen 2008). However, when used by trained embryologists, the developed time-lapse ESAs have shown promising results only in identifying embryos with low developmental potential. Adding TLI systems to conventional manual embryo testing did not improve clinical outcomes but increased embryo morphology assessment times (Chen 2017, Conaghan 2013, Kaser 2016, Kirkegaard 2015). Dimitriadis and colleagues demonstrated a fast and simple cohort embryo selection (CES) method for selecting cleavage-stage embryos that will develop into high-quality BLs. This study demonstrated the ability of embryologists to quickly identify high-quality cleavage-stage embryos when all embryos in the cohort were simultaneously compared in a single image. This selection method outperformed traditional cleavage-stage embryo ranking methods based on morphology and adjunctive morphokinetic TLI parameters (Dimitriadis 2017). This method is excellent at identifying high-quality embryos from a cohort; however, this method of selection is subjective and lacks consistency between operators.

Vision learning technology has been proposed to overcome the labour constraints and subjective nature of assessing and selecting embryos based on morphology and morphokinetic measurements. Kanakasabapathy and colleagues used deep learning CNNs to train and validate embryo assessments on day 3 embryo images based on embryo developmental outcomes recorded on day 5 of culture. This algorithm was trained to make the following day-5 developmental predictions: embryo arrest, morula, early BL, full BL, and high-quality BL. Using a test set of 748 embryos, the algorithm's accuracy in predicting BL development at 70 hpi was 71.87% (CI: 68.41% to 75.15%) (Kanakasabapathy 2020b).

To evaluate the potential improvement in predictive power, Kanakasabapathy and colleagues also compared the accuracy of predictions by embryologists in identifying embryos that will eventually develop into BLs when presented with embryo morphology imaged on days 2 and 3 of development. Additionally, their performance was evaluated with and without the use of Eeva's three-category TLI algorithm that uses P2 (duration of the 2-cell stage) and P3 (duration of the three-cell stage) to predict BL development (VerMilyea 2014). The neural network significantly outperformed the embryologists in identifying embryos that will develop into blastocysts correctly ($P < 0.0001$) and the

overall accuracy in prediction, regardless of the evaluated methodology (P <0.0001). This was the first AI-based system for predicting the developmental fate of cleavage-stage embryos (Kanakasabapathy 2020b).

Bormann and colleagues described an early warning system for using cleavage-stage embryos and statistical process controls for detecting clinically relevant shifts due to laboratory conditions (Bormann 2021). This study presented a novel key performance indicator (KPI) for monitoring embryo culture conditions at the cleavage stage of development. This AI-based KPI predicted the percentage of cleavage-stage embryos that would develop into high-quality blastocysts on day 5 of development. Compared with 5 established cleavage-stage KPIs, this AI-based KPI for predicting high-quality blastocyst formation had the highest association with ongoing pregnancy rates (R^2=0.906). This is the first AI-based cleavage-stage KPI demonstrated to detect changes in a culture environment that resulted in a shift in pregnancy outcomes.

Carrasco et al. used 800 cleavage-stage embryo images with decision tree methods and statistical analysis of features to determine the implantation potential of cleavage-stage embryos (Carrasco 2017). Wang et al. extracted features from textures from 206 micrographs of early embryos (2h of development) (Wang 2018). SVM was used (10-fold cross validation) to achieve 77.67% accuracy and 0.78 of AUC to predict the early embryo development stage (initial and days 1, 2, 3, and 4). Kelly and colleagues used CNNs to identify safe regions on a cleavage-stage embryo to perform laser-assisted hatching. This study utilised more than 13,000 annotated images of cleavage-stage embryos to develop an algorithm identifying the largest perivitelline space region or atretic/ fragmented blastomeres. These regions of the cleavage-stage embryos were considered the safest at which to perform laser-assisted hatching. The AI-trained network was tested on almost 4,000 cleavage-stage images and had 99.4% accuracy with a 95% CI ranging between 99.1% and 99.6% (Kelly 2020). Using CNNs, Meyer and colleagues were able to classify cleavage-stage embryos as aneuploid or euploid with a high specificity and thus successfully identify 85.5% of aneuploid embryos (Meyer 2020). These results demonstrate the ability of CNNs to identify noninvasive markers for detecting genetically abnormal embryos. Collectively, these studies show that various AI techniques can be utilised to extract unique features from cleavage-stage embryos, which may be used for classification, assessment ranking, or to aid in clinic decision-making.

1.3.2.3. AI assessment of blastocysts

A key question about blastocyst assessment needs to be answered: when do we evaluate blastocysts? Since blastocyst development is dynamic, do we evaluate and grade blastocysts when they exhibit the “best” appearance? Or should we evaluate them at a particular time? This question has yet to be answered by existing AI applications, which have utilised both fixed and flexible time-based evaluation methods (Glastein, 2023).

Another issue with blastocyst assessment involves grading. For instance, the problem with using Gartner-type blastocyst grading to assess embryo quality is that it is subjective and does not include quantitative parameters. It is a visual estimate of the number, size, and morphology of the inner cell mass (ICM) and TE cells. On the other hand, BL expansion can be easier to standardise if we use measurement tools and volume ratios. The number and compaction of the cells estimate the quality of the ICM (Gartner 2000). However, the minimum number of ICM cells necessary to develop into a viable human foetus is unknown. In addition, the ICM is a cocktail of pluripotent (epiblast) and primitive endoderm (hypoblast) cells. The size of the ICM alone does not indicate the composition of the cells within.

Assessing TE cells is more challenging, as the cell number, shape, nuclear content, and position in the expanding BL are not standardised. AI methods that use segmentation of the BL will enable us to score TE complement objectively. Judging the ICM's compaction is more accessible than assessing TE quality.

Both blastocyst cell types (ICM and TE) are required for successful implantation. Since current BL grading systems are very simple, it is no surprise that they are not very informative when used to predict implantation. More complex and detailed BL grading systems correlate well with implantation potential and ploidy assessment. In their recent paper, Zhan et al. converted alphanumeric blastocyst grades into a numeric score for statistical analysis and correlations (Zhan 2020a). By using AI, we might be able to strengthen the correlation between blastocyst assessment and outcome more objectively. Also, the ability of AI blastocyst applications predicted by early developmental versus later developmental events needs to be explored.

1.3.2.4. Training databases

Most training data sets used in AI protocols are labelled data, i.e., supervised learning. Humans perform labelling, and thus, it is very subjective. In addition, if clinical outcome data are used,

humans select the embryos for transfer. The requirement for heterogeneous, diverse training data, including an ethnically and racially diverse population of patients, is essential. A balanced data set is also important to eliminate bias in AI learning (Swain 2020). Unsupervised learning is an attractive alternative that needs to be explored.

1.3.2.5. AI and Time-lapse microscopy image analysis

AI algorithms can be applied to “raw” TLI images. Supervised AI training using previously labelled images was developed in a recently described image analysis system (Tran 2019). The labels included BL and morphokinetic annotations with positive or negative implantation results. One of the system's drawbacks was its reliance on humans to create the labels, introducing biased observations and scores. The other problematic practice was using non-viable, non-fertilised, or discarded material for negative training groups to increase the training data set. The rationale behind this was the establishment of a completely automatic system that would also be able to recognise these negative embryos. The question remains: will the developed algorithms perform equally well after removing the discarded group? And are they superior to the BL grading system (Kan-Tor 2020b)?

In another recent study, a different approach was used to predict BL development. It used TLI data up to day 3 of embryo development. Two AI algorithms were developed: an automatic morphokinetic data model (temporal) and a TLI embryo image model (spatial). Both models have comparable predictive power (~0.7). When combined, the different weights were used to optimise blastocyst prediction. Interestingly, more weights were given to the morphokinetic data compared to the images. Compared to embryologists, the AI model performed better in sensitivity and specificity (Liao 2021). In another TLI study, BL prediction was accomplished using morphokinetic TLI data from the first three days of development. Interestingly, by applying a self-improvement (reinforcement) strategy, the predictive power of the AI system improved (d’Estaing 2021).

We have yet to fully discover the oocyte, zygotic, and embryonic development parameters that predict embryo blastocyst development. Early parameters of zygotic (cytoplasmic movement) development, analysed by AI-powered methods, have been shown to be predictive of blastocyst development. Compared to human evaluation and prediction using morphological parameters, AI-based methods using cytoplasmic kinetics showed an average of 10% higher accuracy (Coticchio 2021).

One novel approach to assessing blastocyst quality is using AI to evaluate a quantitative standard expansion assay (qSEA). This measures the kinetics of blastocyst expansion and correlates to the outcome, where faster-expanding blastocysts exhibit higher implantation potential (Huang 2021).

The following novel embryo parameters have been proposed by Bori et al. to be included in AI selection models: pronuclear kinetics, blastocyst measurements, the size of the ICM, and the cell cycle length of the TE cells. The authors' algorithm must be evaluated on the IVF patient population to verify the general utilisation of their proposed model (donor oocytes). The same group presented a novel model utilising AI to predict embryo implantation. Utilising AI image analysis combined with the embryo proteomic profile of PGT euploid embryo spent culture media. The authors were able to demonstrate very high implantation prediction. Although the study is preliminary, it demonstrates the power of AI to combine different data points (proteins and morphology) (Bori 2020).

1.3.2.6. Single-image analysis

The object of a study by Khosravi et al. was to establish an AI deep learning model that can evaluate BL quality (Khosravi P 2019). In this AI-based prediction model, the BL expansion was an important parameter, followed by ICM and TE quality. The precise time point used for the AI evaluation (110 hr) demonstrated the importance of embryo developmental kinetics for embryo prediction. In a 2020 study by Bormann et al., a single image from the TLM image pool at 113 hr was used for analysis (Bormann 2020). A CNN system was used to classify BLs based on the presence of the cavity and the morphological quality of the ICM and TE. Similar to Khosravi et al., Bormann’s group demonstrated that the accuracy of this system for classifying blastocysts versus non-blastocysts was very high (91%). Using the genetic algorithm, the authors established a blastocyst (BL) ranking system called the “BL score. ”The evaluation of the AI BL selection method, using implantation outcomes of the blastocysts selected by humans for transfer, showed over 50% per cent positive outcomes. It will be necessary to perform a comparative prospective study to identify the (dis)agreement in blastocyst selection for transfer between AI models and embryologists. The emerging question is how different the blastocyst selection for ET is between embryologists using the blastocyst grading system (Gartner) and AI model selection. How often will the AI choose a different blastocyst for ET than the embryologist within the cohort of available embryos? There is a lot of disagreement among embryologists grading blastocysts, but how often is the best BL chosen for ET? There are no standards for choosing an AI system for embryo evaluation. They depend on the type of

data, the size of the data set, and the output queries (Fernandez 2020). It will be helpful to compare multiple AI models on the same data set.

Other AI models do not use a specific time point for image analysis. In the model by VerMilyea et al., the “viability” of the embryos was categorised based on the embryologist-given grade (Gartner), where a “3BB” blastocyst was a cut-off for viable and non-viable classes using foetal heart measurements (VerMilyea 2020). Using computer vision image processing and deep learning, the authors achieved an overall accuracy of over 60% and an average accuracy improvement of 24% over embryologist grading.

1.3.2.7. Automatic annotation

One potentially confounding factor that can affect AI protocols is that the morphokinetic annotations are done by humans and are subjective. Developing AI models to recognise abnormal karyokinetic (nuclear) and cytokinetic abnormalities (direct divisions 1–3, cell fusion) will be necessary for optimal automatic annotation.

Most ML embryo assessment and selection methods have used “computer vision methods” utilising visual data (TLI or microscopic images). CNN is a method of choice to process visual information. It can be used for automatic cell annotation (Malmsten 2020), cell detection and tracking (Leahy 2020), embryo grading and selection, and BL and implantation prediction (Louis 2021). All of these studies were done on retrospective data under experimental settings. The clinical application of AI still requires prospective studies.

1.3.2.8. Implantation prediction

In studies using AI to predict embryo implantation potential on static (Chavez-Badiola 2020a, see chapter 3 of this thesis) or TLI images, secondary factors such as laboratory conditions or other human factors have not been analysed or included in the models. Culture conditions and human expertise are important factors that influence embryo development and quality. These factors will need to be included in the models to achieve a useful and objective prediction. In addition, we know that successful implantation and live birth depend on other factors not inherent to the embryo. Predicting implantation solely on embryo quality is an incomplete assessment. AI embryo prediction models should focus on ranking the embryos within the patient cohort rather than on implantation

prediction. The variation in success rates among IVF centres and labs prevents the establishment of universal AI models for implantation prediction (Zaninovic 2020).

1.3.2.9. AI for detecting aneuploidy

As mentioned in Section 1, PGT-A remains the most objective way to assess an embryo. However, its invasive nature, cost, and the sometimes misplaced assumption of diagnostic accuracy limit a more widespread use. It is no surprise that noninvasive approaches for embryo selection, including time-lapse morphokinetic evaluation (Campbell 2013), morphology assessment (Capalbo 2014, Zhan 2020b), and AI systems (Pannetta 2018), have aimed to compare PGT-a outcomes against their findings. However, it is still difficult to find studies presenting AI systems for embryo ranking trained against ploidy status as ground truth. The study presented in Chapter 4 of this thesis is likely the first published work of its kind, demonstrating the potential of using a single static BL image to predict embryo euploidy and assist embryologists in noninvasive embryo selection (for more details, refer to Chapter 4).

We anticipate that other similar full-paper publications will follow shortly, presenting new approaches aimed at embryo selection based on ploidy. These studies will perhaps target time-lapse sequences (Barnes J 2020) and incorporate omics (Bori 2021), noninvasive chromosome screening tests (Chavez-Badiola 2020 c. See Chapter 3), as well as new AI approaches. Building high-quality datasets from diverse settings while managing hype (VerMilyea 2019b) and expectations are challenges that will remain.

Evidence appears to be accumulating in support of using AI for embryo selection. Loewke et al. (2022) conducted a retrospective study evaluating an AI model for ranking blastocyst-stage embryos. The study demonstrated the AI model's potential to improve clinical pregnancy prediction compared to manual morphology grading, although it also identified limitations related to image quality, bias, and granularity of scores that may impact clinical use. Similarly, Salih et al. (2023) presented a systematic review of 20 studies, finding that AI models consistently outperformed embryologists in predicting embryo quality and clinical outcomes. This highlights AI's potential to enhance the reliability of embryo selection in IVF. However, the authors conclude that while the results are promising, further research and validation are necessary before implementing AI for embryo selection in clinical practice.

1.4. Thesis perspectives

AI has long been utilised in other industries and has recently found a place in medical imaging; however, it is just beginning to make an impact on the clinical practice of reproductive medicine, a field familiar to rapid advancements and open to using new technologies to achieve the ultimate goal of a healthy baby.

The IVF process in its current form is still artisanal, subjective, prone to error and difficult to replicate. These flaws impact outcomes and become most evident during decision-making processes, which require highly skilled operators' attention, experience and the ability to interpret images in real-time, as is the case for embryo selection. The introduction of AI has the potential to become a new standard in decision-making, bringing standardisation and replicability in the process by assisting embryologists during decision-making. The challenge remains to develop AI systems that are widely applicable across clinics and independent of differences in protocols and populations. Such is the purpose of this thesis.

1.5. Thesis aims

The specific aims of this thesis of this thesis are:

Specific aim a. To present means of predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning.

Specific aim b. To present the potential of an Embryo Ranking Intelligent Classification Algorithm (termed "ERICA") as an AI-powered clinical assistant with embryo ploidy and implantation predicting capabilities**.** After that, to test the hypothesis that ERICA can be used to predict ploidy status in IVF embryos.

Specific aim c. To use artificial intelligence (AI) embryo selection (ERICA) based on static images in a retrospective pilot study to explore the hypothesis that it can predict first-trimester pregnancy loss.

Specific aim d. To develop the first automatic method for segmenting all morphological structures during the different developmental stages of the blastocyst. After that perform a sensitivity analysis to test the hypothesis that this method is robust.

2. Materials and Methods

Individual methodologies and study design approaches are given in the individual chapters. Unless otherwise specified, ovarian stimulation, oocyte retrieval, embryo culture, preparation for embryo transfer, and embryo transfer were performed according to standard operation procedures and protocols at New Hope Fertility Center clinics in Mexico and are common to all relevant chapters. This chapter describes these steps for reference and is standard practice in IVF.

2.1. Ovarian stimulation and embryo culture

The stimulation, fertilisation, and culture processes followed standard protocols, in brief: following stimulation and oocyte retrieval (Zhang 2016), oocytes were placed in Human Tubal Fluid medium (HTF®, Life Global, EUA) for 2 hours before cumulus stripping. The metaphase II oocytes were inseminated by Intra-Cytoplasmic Sperm Injection and then cultured in a continuous single culture complete medium with Human Serum Albumin (FUJIFILM IrvineScientific, Inc., EUA), covered with Life-Guard Oil ® (Life Global, EUA) to be incubated for 5 to 6 days (CO_2 8.5%, O_2 5%) (Tri-gas; Astec, Japan). The culture medium was renewed after the third day of incubation.

2.2 Embryo transfer

Fresh transfers were performed five days after oocyte retrieval. Luteal support (LS), started on the night of oocyte retrieval, was 400 mg vaginal progesterone (Geslutin, Asofarma, USA) every 12 hours until the pregnancy test and then until week 8 if pregnant. For frozen embryo transfers, endometrial preparation was started with 4 mg oral estradiol (Primogyn, Schering AG, Colombia) from day 3 of the menstrual flow and supplemented with 400 mg vaginal progesterone (Geslutin, Asofarma, USA) every 12 hours starting on cycle day 14–16; both continued until the pregnancy test. If the test was positive, progesterone was continued until week 8, following the same schedule. Blastocysts were thawed 2–4 hours before embryo transfer on day 6 from the beginning of LS. Vitrification and warming were performed using the Cryotop method and vitrification/thawing kits (Kitazato, Japan) as described elsewhere (Kuwayama 2005, Kuwayama 2007).

2.3. Ethical considerations

All patients signed appropriate consents for routine treatment. Since all studies included in this thesis were retrospective for image processing validation with no additional intervention, Institutional Review Board (CONBIOETICA 09-CEI-00120170131) approval was waived (RA-2018-01, 10/2018)

by New Hope Fertility Center's Research and Bioethics Committee. According to Mexican regulation, the above mentioned committee is registered by the National Commission of Bioethics (CONBIOETICA).

All the data employed for the studies included in this thesis was anonymised, and the authors did not have access to any information that could reveal the patient's identities during the study.

3. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning (specific aim a)

The following chapter is encapsulated in a prior published work:

Chavez-Badiola A, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ, Garcia-Sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. Scientific reports. 2020 Mar 10;10(1):4394.

My personal contribution include the original idea of applying AI methods for embryo evaluation, testing different AI approaches, building databases, literature review, analysis of results and discussion.

Some of the specific wording is changed in order to give context to a thesis chapter, however the majority remains unchanged as the words were originally authored by myself.

3.1. Chapter summary

This study addressed the research question “Can positive a beta human chorionic gonadotropin (b-hCG) test after embryo transfer be predicted by assessing both the morphology of embryo and patient features with an automated static digital image analysis system using artificial intelligence and other classifiers?” I, and my co-workers, employed two high-quality embryo micrographs databases with pregnancy outcomes. Together, we created a system consisting of different classifiers that is fed with novel and objective morphometric features extracted from the digital micrographs, along with other non-morphometric data to predict positive h-bCG. The system was evaluated using five different classifiers: a probabilistic Bayesian, a high-dimensional linear classifier (Support Vector Machines), a Deep Neural Network, a decision tree, and Random Forest, using cross-validation to assess the models generalisation capabilities. With the use of the computed morphological features in combination, the SVM classifier and using an objective of 20X, it was possible to achieve an F1 score of 0.76, accuracy of 0.75, and a sensitivity of 0.77. Using an objective of 40X and a Random Forest classifier, a predictive model was created with 0.74 of F1 score, an accuracy of 0.67, and a sensitivity of 0.78. The results obtained indicate the feasibility of using the system to predict positive implantation test from a single digital image to prognosis, offering a potential new and practical approach to embryo classification and selection that may be easily integrated into clinical practice.

3.2. Chapter introduction

As pointed out in Section 1, since the inception of IVF, embryo morphology has been the gold standard for embryo selection. Despite clear and detailed guidelines for morphology-based classifications, embryo morphology assessment remains an artisanal technique lacking objectivity and reproducibility, limiting its overall predictive value. Furthermore, current scoring systems for human blastocysts focus on only 3 parameters, i.e., degree of expansion, quality of inner cell mass and quality of trophectoderm; other features that may have an impact on development potential are essentially disregarded (Yoshida et al., 2018; Sciorio et al., 2018).

The challenges inherent in image-based diagnosis and decision-making are not unique to reproductive medicine. Efforts directed at improving accuracy and standardisation of image analysis through development of computer-aided tools have recently gained attention in other medical fields (Gandomkar et al., 2018; Rebouças Filho et al., 2017; Tang et al., 2017), including dermatology and

oncology, in which linguistic data mining, deep learning techniques and architectures are used for image analysis and to assist diagnosis (Guo et al., 2014; Hu et al., 2018).

The field of reproductive medicine has also ventured into the world of artificial intelligence (AI) (Goodman 2016) but, as yet, with rather limited clinical impact. Evaluating embryo characteristics with digital imaging systems based on quantitative parameters has become an active research area (Lagalla 2015; Rocha 2017a). Efforts have been made to improve objectivity during the embryo selection process by combining multiple morphologic parameters and morphokinetic analysis in mathematical models to improve predictive value for development potential and implantation. Such models, however, are often limited to the initial development stages of the embryo, dependent on currently existing classifications, or require sophisticated time-lapse microscopy systems (Mio 2006, Storr 2015, Majumdar 2017).

3.2.1. Chapter aims

With the above in mind, this first results chapter pursued the following specific aims:

Specific aim ai: To develop a computer-aided classification system, “AIR E®,” designed to identify novel and objective morphologic parameters at the blastocyst stage of development in-vitro, is presented.

Specific aim aii: To test the hypothesis that AIR E® can distinguish implanted *vs* non-implanted embryos, validating using a set of embryos from single embryo transfers.

3.3. Material and Methods

A morphometric analysis of micrographs of blastocysts transferred as single embryos between 2015 and 2019 at two fertility centres in Mexico (dbA and dbB) was undertaken. The databases were tested both independently and combined (dataset named ALL).

A total of 221 blastocysts were photographed on days 5-6 after insemination using either of two inverted microscopes: Olympus IX71 (dbA) or Olympus IX73 (dbB), both with Hoffmann modulation contrast, a digital camera and a Hamilton Thorne ZILOS-tk® Laser camera. Images taken at different magnifications were selected to assess the ability of the computational tool to standardise image processing regardless of magnification: images taken at 20x (total magnification 200x) and images

taken at 40x (total magnification 400x). Each blastocyst was photographed one time only, with a focus on the zona pellucida and trophectoderm.

Feature	dbA	dbB
Objectives (20X, 40X)	(90, 44)	(4, 83)
Donated oocytes (%)	16.92 % *	12.64 %
Mean oocyte age (STD)	34.11 (6.09)	34.93 (4.96)
Embryos expanding	62.69 %	78.17 %
Embryos hatching	32.83 %	17.23 %
Embryos hatched	4.48 %	4.60 %
b-hCG ≥ 20mUI/mL (%)	52.98 %	56.32 %
Proportion of fresh oocytes	13.43 %	

TABLE 3.1 Description of relevant metadata of the databases, including the objectives used, the oocyte source, mean age of the oocyte, the embryo stage, proportion of fresh samples, and b-hCG ≥20. *4 unknown

All transfers were single embryo transfers of D5 blastocysts (≥180u in diameter as measured before transfer), performed following previously published protocols (Zhang et al., 2016). In brief, fresh transfers were performed 5 days following oocyte retrieval. Luteal support was given as 400mg vaginal progesterone (Geslutin, Asofarma, USA) every 12 hours until the pregnancy test. For frozen embryo transfer, endometrial preparation was started with 4mg of oral estradiol (Primogyn, Schering AG, Colombia) from day 3 of the menstrual period and supplemented with 400mg vaginal progesterone (Geslutin, Asofarma, USA) every 12 hours starting on day 14-16 of the cycle; both were continued until the pregnancy test. Blastocysts were thawed 2-4 hours before embryo transfer on day +6 from the beginning of luteal phase support.

Assisted hatching was performed on all embryos. Embryoglue (EmbryoGlue®, Vitrolife, Sweden) was used as transfer medium in all transfers. Transfers were performed with Kitazato catheters (Kitazato, Japan) under transvaginal ultrasound guidance.

3.3.1. Positive b-hCG and confirmation

Seven days following embryo transfer, a quantitative b-hCG assay was performed. A positive pregnancy test was defined as a beta hCG level ≥20mUI/mL.

3.3.2. Image analysis and statistical assessment

An image processing algorithm available in the AIR E® prototype system was used to analyse all 221 blastocysts included in the study (see Figure 3.1.). Using AIR E®, expert embryologists manually delimited the zona pellucida and the trophectoderm. Then, the algorithm converted the images to grey scale and computes 24 morphological features related to perimeters, areas, information quantity, edges, and dispersion of the embryo images. As metadata, we included the age of the patient in years.

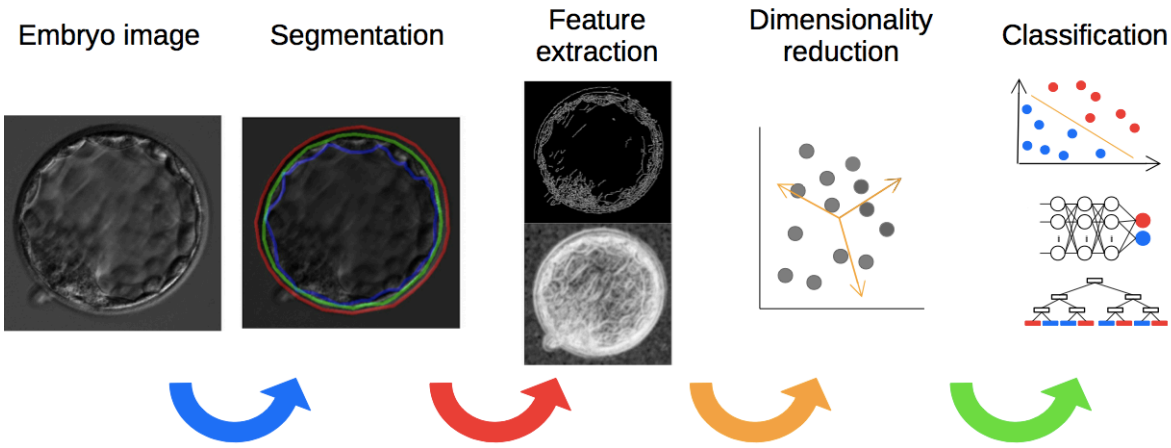


FIGURE 3.1. Overview of the AIR e ® pipeline. Once the embryo image is taken, it is passed through Segmentation, Feature Extraction, Feature Selection, and Classification.

The feature metrics were then converted to micrometres using a factor obtained from each microscope and objective. Afterwards, AIR E® selects the 15 features with greater information gain, which uses the concept of entropy to compute the most useful features. Then AIR E® uses Principal Component Analysis to perform dimensionality reduction. Using the 15 main Principal Components, we trained a supervised model to classify samples with b-hCG ≥ 20mUI/mL. Five models were tested: 1) Naive Bayes, 2) v - Support Vector Machine (v-SVR model for minimisation of the errors, and an RBF kernel with g=0.14), 3) deep Neural Network (three layers of 30 units each and one with 10 units, ReLu activation function, Adam solver, and L2-Regularisation of 0.0001), 4) decision tree (unrestricted), and 5) Random Forest (100 trees). To assess the generalisation capability of the classification models and avoid over-fitting, we used a stratified 10-fold cross-validation, and the reported measures are the average of the results of each fold. We computed the sensitivity, specificity, precision, accuracy, F1 and Area Under the receiver operating characteristic Curve (AUC) to assess the results.

3.4. Results

A total of 221 transferred blastocysts were included in this study: 134 from dbA; and 87 from dbB. Of these, 26.4% (95% CI 20.7% - 32.2%) were hatching blastocysts when the image was taken (2-3 hours post-warming in the case of vitrified embryos), 69.1% (95% CI 62.5% - 75.1%) were fully enclosed by the zona, and 4.5% (95% CI 2.2% - 8.2%) were hatched. The mean maternal/donor age was 34.4 years. Overall positive pregnancy rate was 54.1% (95% CI 47.3% - 60.8%). Table 1 shows the differences and similarities between the two datasets.

3.4.1 Embryo identification capability

Following manual segmentation of images, the mathematical algorithm of the software AIR P.2® was able to extract all features from all 221 images automatically and to compute these for prognosis.

3.4.2. Positive b-hCG prediction

With the 24 morphological characteristics and the age of the oocyte, we extracted the 15 more relevant features for each of the three runs. After this extraction, the 15 principal components were calculated and used for training the four different models. The best models for each database were selected based on the best F1 score: SVMs performed best in dbA and ALL databases, while Random Forest in the dbB database. Figure 2 shows sensitivity, specificity, precision, accuracy, F1 score, and AUC for the best model for the three datasets.

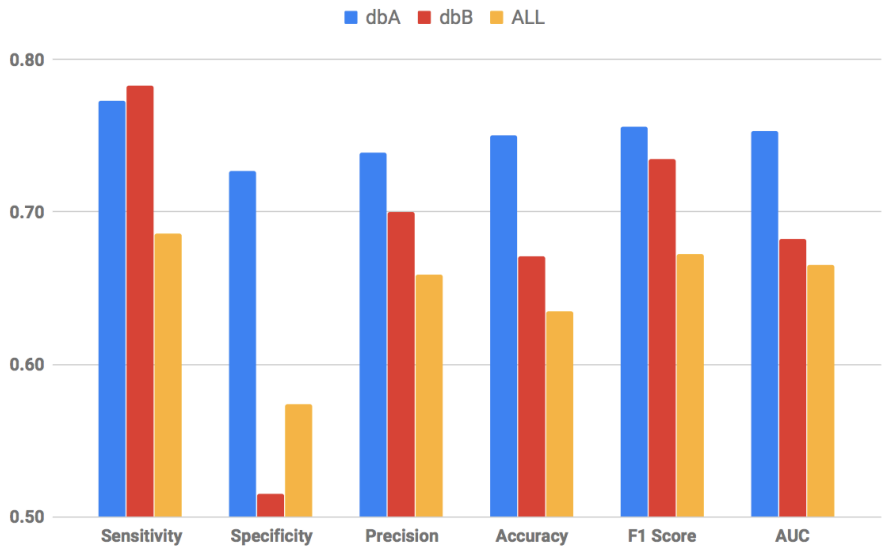


FIGURE 3.2. Comparative chart of the best model for each database. The dbA and ALL databases obtained the best F1 score using the Support Vector Machine model, and the dbB database using the Random Forest model. AUC – Area Under the receiver operating characteristic Curve.

False positive rate (FPR) and false negative rate (FNR) were also computed for all three datasets (See Table 3.2).

	dbA	dbB	ALL
FPR	0.16	0.20	0.19
FNR	0.18	0.13	0.17

TABLE 3.2. False positive (FPR) and negative rates (FNR).

To ensure that the results are not a consequence of over-fitting, we randomised the pregnancy test results over all samples and tested the models. For the dbA database, the SVM obtained an F1 of 0.52; the dbB database obtained an F1 of 0.52 on the RF model, and the ALL database obtained an F1 of 0.57 on the SVM model. Using all the dataset, we obtained precision of 0.66, and sensitivity of 0.69.

3.5. Chapter discussion

The results of this study suggest software prototype AIR E®, is able to extract features that increase the predictions of success in embryo transfer defined as a positive b-hCG test.

In sensitivity, the dbA and dbB databases performed very similar. But specificity and accuracy, dbA overcame dbB because of the high FPR of the latter. The main differences between these two datasets are the objective used in the microscopy and the proportion of expanding/hatching embryos, which suggest that these variables could influence the specificity; however, other confounding factors might influence the outcome.

Using all the dataset, we obtained a sensitivity of 0.69, which could be interpreted as a 69% success rate if the algorithm labels an embryo as a positive b-hCG. This success rate is close to the success rate of an embryo with a euploidy test. However, we emphasise that the present work is only a pilot study, and the validity of the proposed model needs to be shown prospectively with a larger sample size.

Image-based diagnoses and subsequent decision-making processes are challenging in that unless they are objective and standardised, they are likely to impact on prognostic accuracy and efficiency of decision-making. Current embryo selection is in such a predicament since most accepted

classifications are based on subjective assessments during an artisanal process performed by embryologists with various degrees of training and expertise. Furthermore, and perhaps in the interest of simplicity, current non-morphokinetic classifications are structured based on characteristics which require no measurements; they are therefore subjective, disregarding variables or characteristics that cannot be identified by the naked eye, but which could potentially be reflective of development potential.

By contrast, computer-aided image processing tools can identify, in an automated, objective, and replicable fashion, key image characteristics beyond what is recognisable by the “naked eye”. These technologies could thus help transform current classifications from subjective and morphology-based, to objective and standardised image processing. In reproductive medicine, efforts have been directed towards implementing computational tools and artificial intelligence software for the embryo selection process, but until recently (Fishel 2018), most of these efforts have either focused on predicting blastocyst formation (Goodman et al., 2016) or they have been designed to fit current embryo classifications (Rocha 2017b,c).

In a recent study, Rocha et al. (2017a), described a computer-aided image processing tool aimed at identifying morphologic features from blastocysts and then used these features to objectively classify embryos based on current blastocysts scoring systems (i.e. Gartner and Schoolcraft), through a complex neural networks protocol (i.e. genetic algorithm) (Rocha 2017b). Their method still uses existing classification systems, moving this step further through automation but failing to link embryo classification to viability prediction. Our current software prototype, on the other hand, takes a different approach by disregarding existing classifications and linking independent and new variables to prognosis. By bypassing “standard” morphology, the prognosis determined by the algorithm could allow an embryo ranking system and a more straightforward decision-making process during embryo selection for transfer. Another cornerstone of Rocha’s approach is that input information came from micrographs obtained from the EmbryoScope (Rocha et al., 2017a), allowing the investigators to standardise variables related to the use of different imaging protocols and settings (e.g. camera type, objectives, microscope type, etc.). This advantage, however, seems to be tempered by a lack of flexibility, for example, the inability to classify hatching embryos and, even more importantly, a limitation in the form of a requirement for expensive time-lapse equipment. Once again, we decided against a tempting standardised approach but instead trained our algorithm for image extraction and processing after identifying and automatically standardising for imaging variables. In this way, we

aimed to generate a computational tool that could be widely implemented in clinical practice with equipment currently available in most IVF laboratories.

To the best of my knowledge, this is the first time an image processing computational system is used to search for new variables related to prognosis beyond conventionally used parameters found in current classification systems and which lack clear prognostic value. Since classification algorithms thrive on numbers, the authors are currently carrying out a study with a larger sample size and improving the feature extraction algorithm with the final goal of developing a blastocyst ranking system based on prognosis. The validity of the model remains to be demonstrated prospectively.

4. Embryo Ranking Intelligent Classification Algorithm (ERICA), an artificial intelligence clinical assistant with embryo ploidy and implantation predicting capabilities (specific aim b)

The following chapter is encapsulated in a prior published work:

Chavez-Badiola A, Flores-Saiffe A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reproductive BioMedicine Online. 2020 Oct 1;41(4):585-93.

I was awarded the first prize from the Catherine Foundation for Reproductive Medicine during an oral presentation at the 2019 ART World Congress in NYC, presenting results for ERICA's automated embryo segmentation, part of ERICA's algorithm. Previous work cementing the foundations of the algorithm described in this manuscript was also presented during ASRM's 2019 annual meeting as three independent poster presentations and an oral presentation at 'Fertility 2020' in Edinburgh, UK.

My contributions include the original idea of applying AI methods for embryo evaluation, training a CNN using ploidy status as ground truth, building databases, literature review, analysis of results and discussion.

Some of the specific wording is changed in order to give context to a thesis chapter. However, the majority remains unchanged as I originally authored the words.

4.1. Chapter Summary

Selecting embryos with the highest implantation potential remains an imperfect process often based on the subjective assessment of morphology and is prone to inter-observer variability. Recent efforts to improve embryo selection include morphokinetics with time-lapse microscopy and preimplantation genetic screening (with compulsory embryo vitrification/ freezing). In this study, I and my co-workers aimed to test an artificial intelligence (AI) model (ERICA), to rank embryos based on its ability to predict euploidy and pregnancy test results, using a single static blastocyst image in a known outcome data set. Following training and validation, ERICA was more successful than random selection and experienced embryologists in correctly identifying and ranking embryos with the highest implantation potential using a static picture as the only source of information. These encouraging results suggest that ERICA has the potential to assist embryologists and clinicians during embryo selection, without necessarily needing time-lapse incubators or invasive embryo biopsy. In conclusion, the AI model ERICA has the potential to assist embryologists and clinicians during embryo selection, based on ploidy and implantation prediction, without the need for time-lapse or invasive embryo biopsy.

4.2. Chapter introduction

Achieving a successful pregnancy with the minimum of treatment intervention is the aim of every infertile couple and fertility clinic. Unfortunately, factors affecting pregnancy are multifactorial and incompletely understood. Male and female factors can exist, but female age (and subsequent egg quality) is one of the most important confounders. Over 200 confounders exist, affecting outcomes in assisted reproduction (Pool et al., 2012), the majority stemming from the quality of gametes and embryos and the effects of the laboratory and procedures. No single test, observation, or algorithm has been found to predict implantation. Abnormal embryo genetics is the most common cause of miscarriage (Popescu *et al.*, 2018). Other factors such as staff effects, the culture environment, endometrial environment, maternal and paternal health, and ease of embryo transfer will always have some influence on the outcome; therefore, it is improbable that any embryo assessment can currently guarantee 100% prediction of a successful outcome.

Genetic embryo testing technology has been available for some time but has not gained widespread acceptance for several reasons. There is limited evidence that pre-implantation genetic testing for aneuploidy (PGT-A) improves success rates in all patient age groups. Still, there may be a role in

older patients (>35 years), where aneuploidy is more common, thus improving time to pregnancy by not transferring aneuploid embryos (Weissman 2017, Munné 2019). Drawbacks to PGT-a are the invasive nature, cost, requirement to freeze all embryos, high follicular response, and an assumption of diagnostic accuracy of the few cells extracted from the embryo. Future developments include minimally invasive PGT-a, such as blastocoelic fluid assessment, non-invasive PGT-a from analysing spent culture media, and whole-genome assessment. At present, these techniques are still being researched.

Historically, embryologists have performed embryo selection based on the morphological appearances of the embryo at various stages of development. This selection method is the most accepted and commonly used to date. More recently, with the introduction of time-lapse incubators, morpho-kinetic assessment has been introduced in some laboratories, providing a wealth of data points. The majority of laboratories function well without time-lapse incubators, which are expensive and may be time-consuming to operate depending on the depth of assessments.

Artificial Intelligence (AI) is an emerging technology aimed at embryo selection (Chavez-Badiola 2020a –Chapter 1). However, it may also have alternative aims, such as determining the diagnostic accuracy of PGT (Liang 2019). In medical practice, AI is becoming more accepted and promises the ability to analyse vast data sets at speeds not possible for humans to compute in the accepted timeframe of preimplantation culture. The advanced mathematics lends itself well to a static image or video analysis of embryo development and can compute and analyse cryptic morphological features. By nature, it is non-invasive and, in due course, may be affordable and more accurate than established approaches as more data is evaluated and compared between different sources. Embryo grading by conventional morphology evaluation and morphokinetic analysis may not be the best or most accurate way to predict implantation potential, but it is standardised and recognised globally.

4.3. Chapter aims

With the above in mind, the purpose of this chapter breaks down into 3 specific aims:

Specific aim bi: To present an AI model named ERICA (Embryo Ranking Intelligent Classification Algorithm) to rank embryos

Specific aim bii: To present results from training, testing and evaluation for ERICA

Specific aim biii: To present ERICA's potential to rank embryos based on its accuracy in predicting PGT-A (i.e., ploidy) and pregnancy test results.

4.4. Material and Methods

4.4.1. Database description

We interrogated a database of 1231 blastocyst micrographs obtained from a cohort of assisted conception patients treated between January 2015 and June 2019 (Figure 1). The micrographs were taken with one of the following inverted microscopes: Olympus IX71 (laser, 400X, or 200X objectives) (640 X 480 pixels) or Olympus IX73 (400X or 200X objectives) (807 X 603 pixels) using standard light optics. The manually-curated images passed through a series of quality filters: i) sufficient light such that the structures are visible; ii) sharp focus of the zona pellucida and trophectoderm; iii) one embryo per micrograph with no visible instruments and little or no debris in the visual field; iv) showing the entire embryo within the limits of the image (including the zona pellucida); and v) text or symbols in the images should not hinder the visibility of the embryos. All images were taken five or six days after fertilisation before any intervention (i.e., biopsy, cryopreservation, or transfer). The preferred fertilisation technique in the three centres was intra-cytoplasmic sperm injection (ICSI).

A total of 841 embryos had known ploidy status (373 were euploid), and 156 had known beta-human chorionic gonadotropin result (beta-hCG) following a single-embryo transfer. We selected from the database only micrographs with known maternal age and time passed from ICSI to the time of taking the picture referred in hours (Figure 4.1). These conditions resulted in a high-quality database of 840 blastocyst micrographs. An embryo with a good prognosis was defined as either having a report of euploidy after Preimplantation Genetic Testing for Aneuploidy (PGT-A) or a positive beta hCG result, defined as beta hCG $\geq$ 20 mUI/mL (performed on the seventh day following embryo transfer), with priority given to the PGT-a result. This preference is based on unknown factors regarding positive pregnancy tests, such as endometrial receptivity and transfer technique. As a test, PGT-a has probably a lower margin of error. The rest of the images (i.e., aneuploid and negative beta hCG results) were labelled as embryos with "impaired prognosis".

The embryos in the picture database had a mean development age of 130.9 hours (95% CI 130.0-131.8) based on the time of ICSI set at 0 hours. The age range of patients undergoing oocyte extraction was 18-47 years (average 37.1 years, 95% CI 36.5-37.7). 46% of the embryos were labelled as having a "good prognosis" (95% CI 42.5%-49.4%), and 54% as having an "impaired prognosis" (95% CI 50.6%-57.5%). 85% of the embryos were hatching (95% CI 82.4%-87.3%), 4.5%

were hatched (95% CI 3.2%-6.2%), and 10.5% were still within the zona pellucida but expanding (95% CI 8.5%-12.7%).

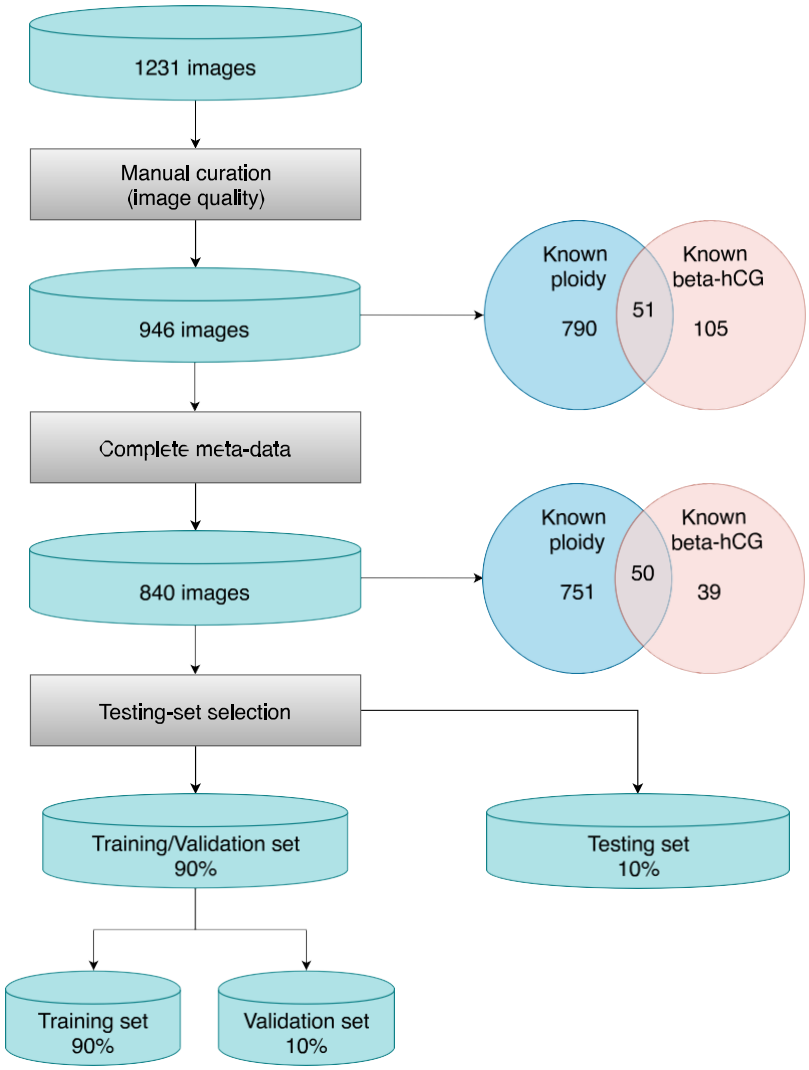


FIGURE 4.1. Database cleaning and setup. Green cylinders correspond to databases, grey rectangles correspond to processes, and Venn diagrams show additional information from a data set.

4.4.2. Training and testing

The full database was split in two: 90% of micrographs were used for the AI model training process, and 10% were used for testing the model in a clinical setting. First, the images to be used for testing the AI model at a later time were separated from the dataset. This testing set (10% of the database) consisted of a group of case scenarios which had never interacted with the model before (never-seen images), comprising the images of a single cycle from 19 patients (84 images). These images were selected from the database using the following inclusion criteria: i) more than one blastocyst available, ii) PGT-a test results available, and iii) having both euploid and aneuploid blastocysts within the same cycle.

The remaining images comprised the training set, which was further divided in two: a “pure training” dataset (80% of the full database), where the algorithm iterates to tune the model and find the parameters that reduce the classification error, and a validation dataset (10% of the full database) exclusively used to assess the performance during each training iteration (also called epoch). Since this set is held back from the training, it could reveal over-fitting on the training set when it is the case. This train/test split approach has proven to produce robust and unbiased models (Vabalas 2019). We define over-fitting as the memorisation of the classification of the samples during the training phase, which is highly undesirable, rather than finding the patterns that explain the outcomes.

4.4.3. Algorithm description

Two modules compose ERICA (beta V.2): The first, designed to extract texture patterns from the micrographs, consists of a pre-processing step on which the images are adjusted to standardise the pixel-to-micrometer ratio. Then, each image is convoluted with 275 custom-designed kernel filters to identify and crop the embryo. Next, each image is automatically segmented into four regions of interest (i.e., background, zona pellucida, trophectoderm, and inner area). Finally, a feature extractor designed to quantify predictors of embryo viability (e.g. size, shape of the inner mass, total energy and entropy measurements) is used to generate a feature vector of size 94 for each micrograph.

The second module is designed to rank embryos based on the identification and scoring of blastocysts. For this, ERICA uses the extracted image-based features and combines them with the metadata of each embryo. This second module consists of a binary classification model generated by a deep neural network (DNN) built on Python’s framework and the above-described training set. The DNN architecture is constructed over three layers of 100 fully connected units, a dropout of 0.4 and L1/L2 regularisation of 0.00001, followed by a two-unit softmax layer. The training was performed through 1000 epochs with an early stopping of 100 epochs of patience with a minimal delta of 0.001 in the validation loss, Adam optimiser with a learning rate of 0.001, categorical cross entropy as the loss function, and 10% of the training set as validation.

The resulting model classification and confidence allow us to define a score for each blastocyst to have a good prognosis. Finally, ERICA employs this score to rank the embryos of a given cycle so that those with the best scores are at the top. Figure 2 depicts a scheme of ERICA´s system parts.

This work employed several open-source libraries (i.e., Open-CV, Skimage, TensorFlow, and Keras).

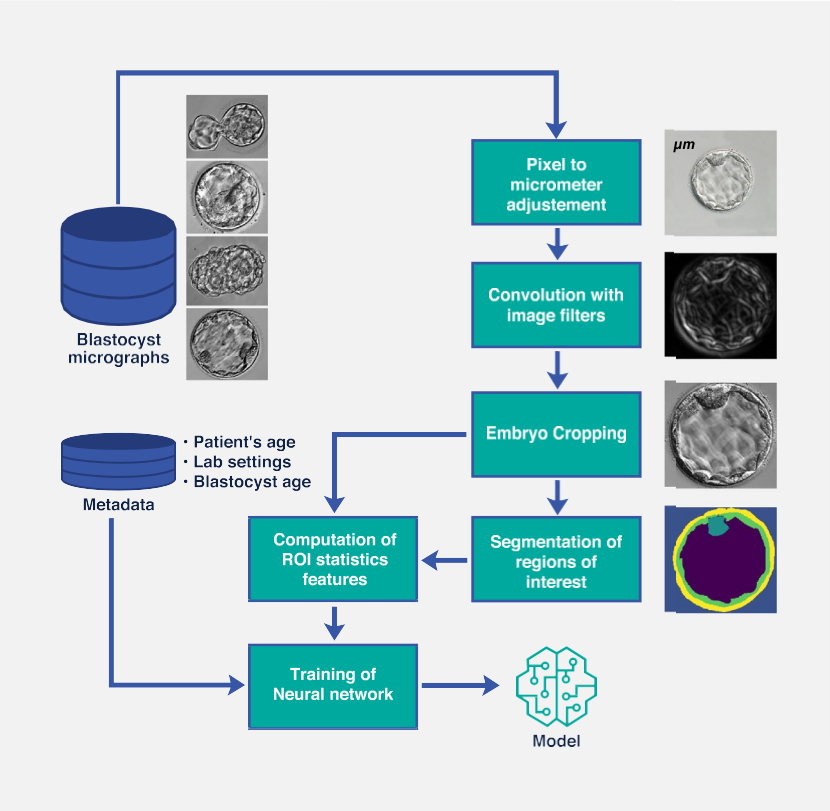


FIGURE 4.2. Algorithm description. ROI, region of interest.

4.4.4. Evaluating the algorithm

ERICA was assessed using four approaches: i) using the testing set to predict euploidy; ii) comparing its ploidy predictions against randomly assigned labels (“good” and “bad” prognosis); iii) comparing its ploidy predictions against those of two senior embryologists analysing still pictures of embryos; and iv) quantifying its ability to select a euploid embryo at the first position of the ranking of the 19 cycles.

To assess the ploidy predictions, ERICA was compared against chance (randomly assigned ploidy predictions) and two embryologists using a metric called normalised discontinued cumulative gain (NDCG). This metric measures the quality of a ranking using a weighted relevance scale based on the position of the list (Järvelin and Kekäläinen, 2002). NDCG scores 1 when it is a perfect ranking (all euploid at the top) and scores lower otherwise. To compute the random ranking, we i) randomly assigned the ploidy status within each cycle and ii) computed the NDCG. This process was repeated 100 times and obtained the mean of the obtained NDCGs for each cycle (Figure 3). The two senior embryologists were asked to rank the embryos within each of the 19 cycles by looking at the picture (the limitation of ranking from a single still picture is discussed below) and considering information

provided for maternal age and embryos' age, defined as time passed from fertilisation to reaching blastocyst status (time from fertilisation to picture). They were asked to rank the embryos within each cycle using numbers from 1 to "*n*," where 1 is the blastocyst with the best prognostic, and "*n*" is the number of blastocysts of each cycle. To calculate *p* values, we conducted Mann-Whitney U tests.

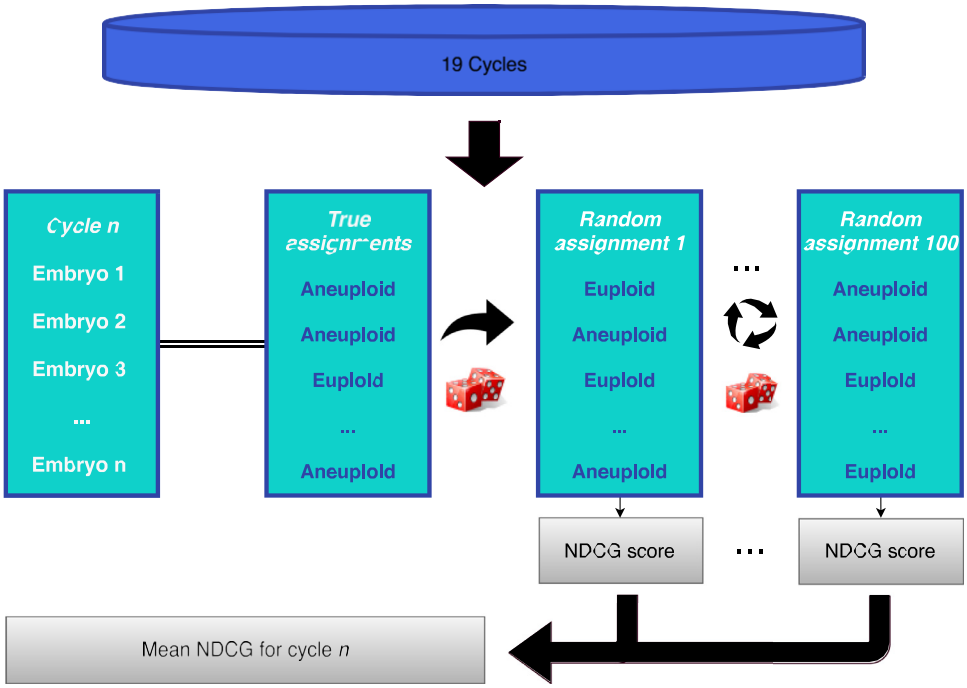


FIGURE 4.3 Process of random analysis. The blue cylinder corresponds to the testing set, black arrows represent processes, and turquoise rectangles are examples of true and random assignments of the known ploidy status for each embryo. NDCG, normalised discounted cumulative gain.

The most senior embryologist (i.e., laboratory director) from two of the three participating centres were selected to classify embryos. The laboratory director for the third IVF centre could not participate in the classification process.

4.4.5. Ethical considerations

All patients signed appropriate consents for routine treatment. Since this study was a retrospective study for image processing validation with no additional intervention, and no embryos were selected for transfer by the AI system, IRB approval was waived.

4.5. Results

4.5.1. Training ERICA

For the prognosis classifier of ERICA, collected a data set of 840 high-quality blastocyst micrographs was selected. The embryos in the picture database had a mean development age of 130.9 hours (95% CI 130.0-131.8) based on the time of ICSI set at 0 hours. The average age at oocyte extraction was 37.1 years (95% CI 36.5-37.7). 46% of the embryos were labelled as "good prognosis" (95% CI 42.5%-49.4%), and 54% as "impaired prognosis" (95% CI 50.6%-57.5%). 85% of the embryos were hatching (95% CI 82.4%-87.3%), 4.5% were hatched (95% CI 3.2%-6.2%), and 10.5% were still within the zona pellucida but expanding (95% CI 8.5%-12.7%).

Using deep neural networks and artificial vision, ERICA extracted 94 features from all the samples. These features were introduced into the training software, achieving a 0.70 accuracy in the validation set.

4.5.2. Testing ERICA

The testing dataset consisted of 84 images from 19 cycles with a mean of 4.4 embryos per cycle (95% CI 3.60-5.24), from which 48.8% were euploid embryos (95% CI 37.7%-60%). On this testing set, ERICA obtained an Accuracy of 0.70 with a positive predictive value of 0.79 and an Area Under the Receiver Operating characteristic Curve (AUC) of 0.74 for predicting euploidy (Figure 4.4). Additionally, ERICA presents a sensitivity of 0.54 and a specificity of 0.86. Table 4.1 shows the confusion matrix (also known as the error matrix) to quantify the type of errors (and hits) of the model, which displays the true and false positives and negatives.

	True euploid	True aneuploid
Predicted euploid	22	6
Predicted aneuploid	19	37

Results represent predictions made by Embryo Ranking Intelligent Classification Algorithm for ploidy status in the test set.

Table 4.1. Confusion matrix of predicting euploidy in a test set

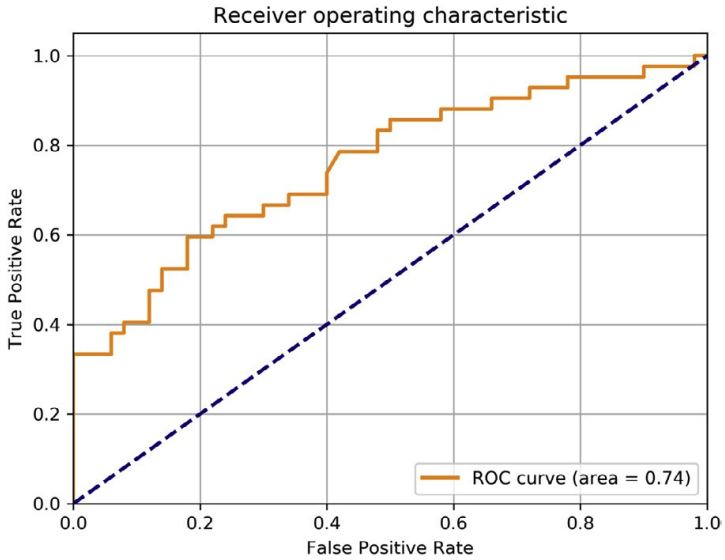


FIGURE 4.4 Algorithm's performance: receiver operating characteristic (ROC) curve. The ROC curve depicts the relation between the true positive rate (y-axis) and false positive rate (x-axis) of the testing set.

After training, the ability of ERICA to rank the embryos against chance (random ranking) was contrasted by two embryologists from different fertility centres using the NDCG metric (see materials and methods). Figure 4.5 shows the NDCG scores for random selection, the two embryologists, and ERICA. These results show that ERICA had greater NDCGs than random selection ($U = 289$, $p = 0.0007$), embryologist 1 ($U = 254.5$, $p = 0.0014$); and embryologist 2 ($U = 246.5$, $p = 0.0242$). One embryologist (#2) could not rank two of the cycles due to the collapsed state of the blastocysts as opposed to ERICA, which achieved to rank all cycles.

Finally, it was established if ERICA could i) find a euploid blastocyst at the top of the ranking or ii) find at least one euploid embryo within the top two blastocysts. The algorithm resulted in 78.9% (15 out of 19) and 94.7% (18 out of 19), respectively, higher than random classification and the two embryologists, as can be observed in (Table 4.2).

	Euploid at the top (%)	**Euploid at the top two (%)**
Aleatory	50.5	84.3
Senior embryologist 1	63.1	84.2
Senior embryologist 2	70.7	94.7
ERICA	78.9	94.7

ERICA, Embryo Ranking Intelligent Classification Algorithm.

TABLE 4.2. Finding a euploid blastocyst at the top of the ranking

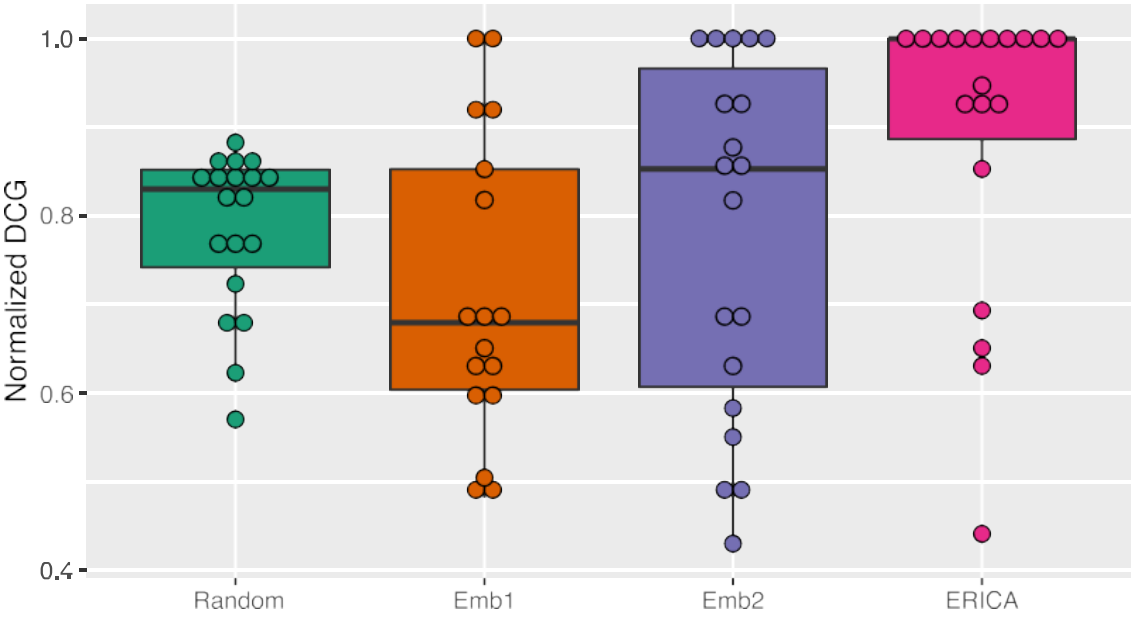


FIGURE 4.5 Ranking results. The normalised discounted cumulative gain (NDCG) is presented for random, embryologists 1 and 2 (Emb1, Emb2) and ERICA. An NDCG value of 1 is interpreted as a perfect ranking where euploid blastocysts were ranked at the top and aneuploid ones at the bottom. The NDCG performance of ERICA is statistically higher than all others.

4.6. Discussion

AI has overcome chance and human performance in several clinical applications not only because of its ability to quickly learn from big datasets (e.g., medical images), but also because it can weigh the relevance of the variables (high precision variable weighting) and their combinations, and detect complex (non-linear) patterns. These AI qualities could aid in improving the current embryo grading system used by embryologists, which could be non-reproducible and sub-optimal within the clinical practice (Adolfsson and Andershed, 2018). As an example of AI's potential, in this study, ERICA successfully extracted 94 features using deep neural networks and artificial vision from all of the pictures in the database, in contrast to the 3 standard features assessed by embryologists to classify according to the most popular embryo grading system (Gartner DK, Schoolcraft WB., 1999), potentially giving ERICA the edge when ranking based on still images.

An ideal AI for this purpose should be trained on objective, reproducible parameters such as PGT-a or a pregnancy outcome (implantation, presence of sacs, heartbeat, or live birth). This way, the AI would learn over a standardised and reproducible procedure, rather than on subjective observations. With this approach, and in addition to relevant clinical data from each patient and cycle, an AI could be comparable to or improve the embryologist's performance. This study presents a robust dataset to

train and test ERICA reliably. For this, we focused on including only high-quality micrographs with known and objective outcomes (i.e., ploidy and beta hCG results). Furthermore, we procured a well-balanced dataset with almost half of the embryos being euploid (46% in the training set and 48% in the testing set), allowing, on the one hand, for reliable training of the algorithm and, on the other, for trustable results from the test.

Tran et al. (2019) recently proposed an automated AI approach using time-lapse videos to predict pregnancy with foetal heart (FH) and obtained an impressive 0.93 of the AUC (Tran 2019). In more detail, their proposed algorithm to predict pregnancy following multiple embryo transfers obtained a true negative rate of 94.5% and a true positive rate of 45.9%. However, these results must be carefully evaluated because they failed to depart from a well-balanced dataset, which could, in turn, reduce the algorithm's reliability (Lobo 2008, Saito and Rehmsmeier 2015). Because the dataset was unbalanced (negative FH accounted for 92% of their full dataset), their true negative rate highly resembles the negative FH in their dataset, risking biased training of their algorithm. Another desirable quality of an AI assisting system is its flexibility when applied to varied laboratory settings, such as the make and age of the microscope, its settings, and the quality of the lens. Recently proposed AI systems (Khosravi 2019, Tran 2019) rely exclusively on morpho-kinetic data extracted from standardised time-lapse microscopes, which is a highly expensive piece of equipment that only a proportion of clinics can afford and whose advantages remain to be conclusively proven (Kieslinger 2016, Armstrong 2019). In this study, we used data from three clinics and five different laboratory settings (two laboratories used two different microscope presets). Despite such differences, ERICA obtained a euploid (positive) predictive value of 0.79 and an aneuploid (negative) predictive value of 0.66, with no differences in predictive values found amongst clinics or settings. This flexibility to adapt to different instruments, we believe, is a significant first step towards having an AI system that could be used in all laboratory settings.

Are these predictive values acceptable? External factors influence prognosis beyond the embryos' potential and cannot be assessed in the clinic before the transfer (at least with current technology). How much do such factors influence chances: 20-25%? Depending on these effects, we believe that a 70% accuracy is currently acceptable. With time, ERICA will be further trained to learn and should be able to improve.

Despite the low sensitivity described in this study, which means a high false aneuploid prediction rate, our results suggest that ERICA outperformed chance (random classification) and classification by two embryologists when they were asked to rank the embryos based on a single static picture from 19 cycles when blinded to ploidy status. ERICA also selected a euploid blastocyst in the first position in 15 of the 19 cycles analysed, improving on our embryologists and randomly labelled results. The latter evaluation was designed to have a clinical perspective, comparing embryologists grading with our algorithm (Figure 6). Another advantage of ERICA is that it includes any blastocyst stage: it can analyse expanding, hatching, or hatched embryos, a feature that, to our knowledge, no other algorithm has. Although this study was not designed to test for speed, following feature extraction, ranking time per embryo approached 25 seconds, which allowed ERICA to yield results in just over a minute (an average of 4.4 embryos per cycle in the testing set). The process could be accelerated, but our goal remains to improve accuracy. Also, it is relevant to highlight that ERICA was trained with data from three clinics only and with a limited dataset size. This limitation is relevant since its accuracy needs to be further tested on other clinics and the database increased.

A limitation of the algorithm is that it is not trained to rank day 3 embryos but only blastocysts. There is ongoing discussion over the advantages and disadvantages of transferring day 3 or day 5-6 embryos (Green 2016, Hatırnaz and Kanat Pektaş 2017). Also, we acknowledge that in this study, random and embryologists' classification of embryos from micrographs appeared comparable. Still, it is based on a small sample of embryologists and restricting the information presented to them. In practice, embryologists who decide to transfer an embryo based on morphology alone do this depending on what they were taught. This involves watching the embryos on a monitor or through a microscope in most laboratories. In other words, it is not a still but a live image. They also have historical information about the oocytes, zygotes, and embryo development if they observe embryos daily or every other day. Additionally, some embryologists may roll the embryos around with a probe to obtain additional information and would not be content to base their decision on a single still image.

The use of ERICA could be advantageous in cycles with unknown PGT-A or where the current grading system is inconclusive. Many clinics offer PGT-a as an additional invasive screening test to increase the likelihood of pregnancy. However, recent evidence suggests that invasive genetic testing is a risk factor for preeclampsia (OR = 3.02) and placenta previa (OR = 4.56) (Zhang et al., 2019). The risks of PGT-a are still controversial (Patounakis and Hill, 2019). Moreover, the efficacy of

PGT-a remains under scrutiny, as recent randomised trials have shown (Munne 2017, Munné 2019, Verpoest 2018). Importantly, an effective AI should be able to select a single embryo (either on day 3 or day 5) with the best chance to improve the time to pregnancy by avoiding multiple sequential transfers or egg retrievals and to achieve pregnancy by avoiding unnecessary risks such as multiple pregnancies and potentially reducing the unnecessary upset of miscarriage.

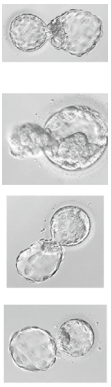
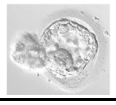
Original embryo ID[a]	Micrograph	Days to reach blastocyst	Embryo grading[b]	Ranking[c]	Classifier result[d]	Ploidy status
E4		5	5,2,1	C	0.93	Euploid
E5		6	5,1,1	B	0.94	Euploid

FIGURE 4.6 Example of a cycle extracted from the testing set. Note the embryo grading assigned by embryologists (b); and the ranking assigned by Embryo Ranking Intelligent Classification Algorithm (c) based on the classifier results (d), and the actual result on the preimplantation genetic testing for aneuploidy report.

The use of ERICA could be advantageous in cycles with unknown PGT-A or where the current grading system is inconclusive. Many clinics offer PGT-a as an additional invasive screening test to increase the likelihood of pregnancy. However, recent evidence suggests that invasive genetic testing is a risk factor for preeclampsia (OR = 3.02) and placenta previa (OR = 4.56) (Zhang et al., 2019). The risks of PGT-a are still controversial (Patounakis and Hill, 2019). Moreover, the efficacy of PGT-a remains under scrutiny, as recent randomised trials have shown (Munne 2017, Munné 2019, Verpoest 2018). Importantly, an effective AI should be able to select a single embryo (either on day 3 or day 5) with the best chance to improve the time to pregnancy by avoiding multiple sequential transfers or egg retrievals and to achieve pregnancy by avoiding unnecessary risks such as multiple pregnancies and potentially reducing the unnecessary upset of miscarriage.

This study demonstrated the potential ability of an AI model, ERICA, to successfully rank blastocysts based on its accuracy in predicting PGT-A (i.e., ploidy) and pregnancy test results. ERICA was more

successful than random selection and senior embryologists in identifying and ranking blastocysts with the greatest potential on the basis of observing one static picture. Although further studies are required, these encouraging results suggest ERICA's potential to assist Embryologists in the embryo selection process without the absolute need for time-lapse incubators or invasive PGT-A. These results support the design of a prospective study where ERICA would be tested with a new data set, with different laboratory culture conditions and microscopes, to confirm pattern recognition and feature extraction reproducibility. Evaluating mosaic embryos is an important topic, as is the cost-effectiveness of ERICA alone or in combination with other tests such as non-invasive PGT-A or metabolomics versus no testing versus invasive PGT-A in different health settings. All these have either started or are planned.

5. Use of artificial intelligence (AI) embryo selection (ERICA) based on static images to predict first-trimester pregnancy loss: a retrospective pilot study (specific aim c)

The following chapter is encapsulated in a manuscript accepted for publication:

Chavez-Badiola A, Flores-Saiffe Farías, Gerardo Mendizabal-Ruiz, Silvestri G, Griffin DK, Valencia-Murillo R, Andrew J. Drakeley AJ, Cohen J. Use of artificial intelligence (AI) embryo selection based on static images to predict first-trimester pregnancy loss: a retrospective pilot study. In preparation for publication.

My contributions include the original idea of linking an AI method for embryo evaluation trained on ploidy potential and the risk of aneuploidy, building databases, literature review, analysis of results and discussion.

5.1. Chapter Summary

This chapter aimed to test the hypothesis that the artificial intelligence (AI) embryo selection assistant (ERICA) developed in the previous chapter can predict the incidence of first-trimester spontaneous abortion (SA) using static images of IVF embryos. A cohort of 172 blastocysts from IVF cases with single embryo transfer and a positive biochemical pregnancy test was retrospectively ranked by the AI morphometric algorithm ERICA™. Making use of static embryo images, each blastocyst was assigned one of four possible groups (Optimal, Good, Fair, Poor), and linear regression was used to correlate the results with the presence or absence of a normal foetal heart rate (FHR) as an indicator of ongoing pregnancy or SA, respectively. Additional analyses included modelling for recipient age and chromosomal status established by preimplantation genetic testing (PGT-A). Embryos grouped as Optimal/Good had a lower incidence of SA (16.1%) compared to embryos classified as Fair/Poor (25%, P=0.005, Odds ratio (OR)=0.46). The incidence of SA in chromosomally normal embryos was 13.3% for Optimal/Good embryos and 20.0% for Fair/Poor embryos, although this difference was not statistically significant (P=0.531). There was a significant association between embryo rank and recipient age (P=0.018), and the incidence of SA was unexpectedly lower in older recipients (21.3% in ≤35, 17.9% in 36-38, 16.4% in ≥39, P=0.0181, OR=0.354). Overall, these results support a correlation between the risk of SA and embryo rank as determined by AI; the classification accuracy was 67.4%. This preliminary study suggests that AI (ERICA™), designed as a ranking system to assist with embryo transfer decisions, might also help provide information for couples on the risk of SA. Future work should include a larger sample size as well as karyotyping of miscarried pregnancy tissue.

5.2. Chapter introduction

Spontaneous abortion (SA), otherwise known as early miscarriage or pregnancy loss in the first trimester, is a common complication of pregnancy (Kolte 2015). It accounts for up to 25% of naturally conceived pregnancies and around 11% of IVF cases. SA can be a devastating experience for patients, leading to a range of stressful emotions and grief, with many couples giving up on further treatment before they become parents (Rooney 2018, Domar 2018). SA management also lengthens the time to the next opportunity for pregnancy, thus prolonging the infertility journey (Munné 2019). The risk of miscarriage increases with maternal age, with the most common cause being chromosomal abnormality, principally aneuploidy (Cimadomo 2018, Gruhn 2019). Currently, the most common intervention for poor prognosis patients with a high likelihood of SA is preimplantation

genetic testing for aneuploidy (PGT-A). PGT-A is perhaps the most controversial procedure in assisted reproduction technology (ART) because of its invasive nature (embryo biopsy is required), its expense, and mixed messages from randomised controlled trials regarding its ability to improve (cumulative) pregnancy rates.

Artificial intelligence (AI) applications are now being widely used in many areas of clinical medicine and medical research (Curchoe 2020, Yu 2018). ART lends itself particularly well to AI applications since it usually generates treatment-associated data in various formats, including images, video, natural language, and time. These parameters can potentially be used to train machine-learning algorithms to predict treatment success at different stages. Examples of these include AI-enhanced embryo selection (Bori 2021, Chavez-Badiola 2020, Tran 2019, VerMilyea 2020), sperm selection (Zhang 2021), semen analysis (Mahdavi 2011, Hicks 2019), and many others (Curchoe 2020a, Kragh 2021, Curchoe 2020b, Riegler 2021). AI is attractive because it facilitates the interrogation of large amounts of data, far greater than the human eye can see or that the mind can process, in a very rapid time frame, thereby creating the possibility of identifying patterns and correlations amongst features not previously recognised as being important (Bori 2020, Gartner 2000a).

ERICA™ (IVF 2.0 Limited, UK) is an AI system initially trained to anticipate the ploidy potential of blastocysts based on an assessment of either single static images or single images it extracts from time-lapse videos and metadata (Chavez-Badiola A 2020b). ERICA™ first extracts texture patterns from the static images. This comprises a pre-processing step in which the images are adjusted to standardise the pixel-to-micrometre ratio, and each image is then convoluted with kernel filters and segmented into four regions of interest, i.e. background, zona pellucida, trophectoderm and inner cell mass. A feature extractor designed to quantify predictors of embryo viability, *e.g.* size, shape of the inner cell mass, and total energy and entropy measurements, is used to generate a feature vector. Secondly, ERICA™ ranks embryos based on the identification and scoring of blastocysts using extracted image-based features and combining them with the metadata for each embryo using a binary classification model generated by a deep neural network. A total of 94 features have been successfully extracted by ERICA using these deep neural networks and artificial vision. The resulting model classification and confidence define a score for each blastocyst in a given cycle and rank them according to their prognosis. In our previous study (Chavez-Badiola A 2020b), these features were introduced into the training software, achieving a 0.70 accuracy in the validation set and a positive predictive value of 0.79 for euploidy (as determined by PGT-A). The ERICA™ algorithm assigns each

embryo a value between zero and one, based on its putative anticipated euploidy potential, categorising it as either “Optimal”, “Good”, “Fair”, or “Poor”. The purpose is to assist embryologists in ranking the order in which embryos from a given cohort are to be transferred (Chavez-Badiola A 2020b).

5.2.1. Chapter specific aims

Based on the known association between aneuploidy and SA, the aim of this chapter was:

Specific aim ci. To perform an initial feasibility study to establish whether embryo ploidy ranking using ERICA™ could potentially be used to predict the occurrence of SA. Specifically, to test the hypothesis that events occurring between implantation (defined as a positive biochemical pregnancy result) and clinical pregnancy (defined as the detection of a foetal heart) correlated with an Optimal/ Good/Fair/Poor prediction using ERICA™. I chose this precise definition to reduce the impact of underlying confounders, such as the ovarian stimulation regime and the embryo culture system used.

If this hypothesis proves correct, such information could potentially be beneficial when counselling couples (or individuals) regarding their embryo quality.

5.3. Chapter methodology

A retrospective cohort of patients undergoing ART in two Mexican fertility clinics herein investigated; only single embryo transfer cases that proceeded to a biochemical pregnancy were included in the analysis.

5.3.1 Ethical approval, data collection, definitions, and study inclusion criteria

Institutional Review Board approval was granted for this project (CONBIOETICA 09-CEI-00120170131). A database of 525 embryos transferred between August 2019 and May 2021 at two IVF clinics in Mexico (New Hope Fertility Centres, Guadalajara and Mexico City) was interrogated for which images had been uploaded into the ERICA™ web application before their transfer (erica.embryoranking.com).

A positive biochemical pregnancy was defined as a quantitative beta hCG $\geq$20 IU/L seven days post-transfer. Foetal heart rate (FHR) status was defined as either “present” or "absent” based on week

7-8 transvaginal ultrasonography (TVS). SA was defined as either an initial positive pregnancy test (beta hCG ≥20 IU/L) followed by a subsequent fall in beta hCG values; transvaginal bleeding equal to or heavier than a menstrual period after excluding an ectopic pregnancy; the absence of a gestational sac at TVS performed 2-3 weeks after first biochemical pregnancy test; or a pregnancy which failed to develop an FHR identifiable with TVS by 7-8 weeks (3-4 weeks after biochemical pregnancy test). An ongoing pregnancy was defined as the presence of an identifiable FHR during TVS performed at 2-4 weeks following transfer (pregnancy week 6-8).

All single embryo transfer cycles with a positive biochemical pregnancy test, known FHR status (present or absent), and registered in the ERICA web application were included for analysis. Pregnancies with a confirmed SA following a positive biochemical test were included even if no TVS was performed. After exclusion criteria were applied, 172 embryo transfers were suitable for analysis (Fig. 5.1). The biochemical pregnancy rate per single embryo transfer in this cohort was 39.6% (n=248/510).

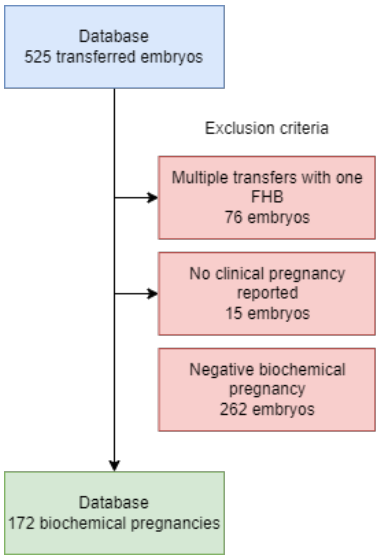


FIGURE 5.1. Number of embryo transfer cycles performed in the period of study (Aug 2019 - May 2021) and detailed inclusion process

5.3.2. Ovarian stimulation and embryo culture

The stimulation, fertilisation, and culture processes followed standard protocols and were described in the general material and methods section. In brief: following stimulation and oocyte retrieval (Zhang 2016), oocytes were placed in Human Tubal Factor medium (HTF®, Life Global, EUA) for 2 hours before cumulus stripping. The metaphase II oocytes were inseminated by Intra-Cytoplasmic Sperm Injection and then cultured in a continuous single culture complete medium with Human

Serum Albumin (FUJIFILM IrvineScientific, Inc., EUA), covered with Life-Guard Oil ® (Life Global, EUA) to be incubated for 5 to 6 days (CO_2 8.5%, O_2 5%) (Tri-gas; Astec, Japan). The culture medium was renewed after the third day of incubation.

5.3.3. Embryo transfer

Fresh transfers were performed five days after oocyte retrieval. Luteal support (LS), started on the night of oocyte retrieval, was 400 mg vaginal progesterone (Geslutin, Asofarma, USA) every 12 hours until the pregnancy test and then until week 8 if pregnant. For frozen embryo transfers, endometrial preparation was started with 4 mg oral estradiol (Primogyn, Schering AG, Colombia) from day 3 of the menstrual flow and supplemented with 400 mg vaginal progesterone (Geslutin, Asofarma, USA) every 12 hours starting on cycle day 14–16; both continued until the pregnancy test. If the test was positive, progesterone was continued until week 8, following the same schedule. Blastocysts were thawed 2–4 hours before embryo transfer on day 6 from the beginning of LS. Vitrification and warming were performed using the Cryotop method and vitrification/thawing kits (Kitazato, Japan) as described elsewhere.

5.3.4 Embryo Imaging and AI-Assisted Ranking

Pictures from blastocyst-stage embryos were taken before any interventions(i.e., biopsy, vitrification, or transfer). For evaluation, Single static images were uploaded into the ERICA™ web app (erica.embryoranking.com). Embryos were photographed with total magnifications of either 200X or 400X, using standard cameras (Hamilton Thorne ZILOS-tk Laser) installed on inverted microscopes (Olympus IX71 or Olympus IX73) equipped with Hoffmann modulation contrast.

Embryos were assigned an output value ranging from 0 to 1 based on their calculated individual ploidy potential. This score was then used to rank all embryos within a given cohort. A technical description of ERICA™ is given elsewhere (Chavez-Badiola 2020b). Based on ERICA™'s output, embryos were arbitrarily defined as follows:

a. Optimal: score ≥ 0.70
b. Good: score ≥ 0.50 and < 0.70
c. Fair: score ≥ 0.30 and <0.50
d. Poor: score < 0.30

5.3.5 Statistical analysis

Based on the hypothesis that embryos with the lowest scores will have higher SA rates, we compared ERICA™'s categories for their potential link to outcomes, defined as the presence or absence of an FHR following an initial positive pregnancy test (biochemical pregnancy). The proportion of SAs within each category was calculated as the total number of biochemical pregnancies with no FHR divided by the total number of biochemical pregnancies. Data are presented as percentages with their calculated 95% Confidence Intervals (CI). Due to the low sample size in some comparisons, embryos assigned to the Optimal and Good groups were pooled, as were embryos assigned to the Fair and Poor groups.

Statistical analysis was completed on SPSS (v.28, IBM) and used a generalised linear regression model using logit link functions for binary outcomes. The model tested for an interaction between embryo ranking and female age since aneuploidy increases, positively correlated with aneuploidy (Hassold 2009). Age brackets of ≤35, 36-38, and ≥39 were chosen arbitrarily. Additionally, a subset of embryos from this database (n=50) had a known euploid diagnosis as detected by high-resolution Next Generation Sequencing PGT-A screening; the effect of this factor was also evaluated.

5.4. Results

The demographics and characteristics of the biochemical pregnancies included in this study are described in Table 5.1. A full breakdown of the entire embryo database is presented in Table 5.2.

Metric	Magnitude
Number of pregnancies included	172 embryos
Donor eggs (%)	28 (16.3%) embryos
Fresh transfers (%)	15 (8.7%) embryos
Known euploid transfers (%)	50 (29.1%) embryos
Early Pregnancy losses (%)	32 (18.6%) embryos
Mean patient's age (STD)	36.2 (5.01) years old

STD – Standard deviation

TABLE 5.1. Demographic information for the study

Groups	Optimal		Good		Fair		Poor	
	n	SA	*n*	SA	*n*	SA	*n*	SA
All	82	16 (20%)	42	4 (10%)	30	6 (20%)	6	18 (34%)
1: (<35)	32	7 (22%)	16	0 (0%)	5	3 (60%)	8	3 (38%)
2: (36-38)	29	4 (14%)	14	2 (14%)	8	1 (13%)	5	3 (60%)
3: (>39)	21	5 (24%)	12	2 (17%)	17	2 (12%)	5	0 (0%)
All known euploids	20	3 (15%)	10	1 (10%)	14	3 (21%)	6	1 (17%)

TABLE 5.2. Total distribution of early pregnancy loss (SA) according to ERICA-assigned embryo ranks (Optimal, Good, Fair, Poor) and age groups. For each subset, the total sample size (n) is given together with the number and calculated %incidence of SA.

5.4.1 Relationship between early pregnancy loss (SA), embryo ranking, and recipient age

When a generalised linear model was employed, statistical analysis identified significant effects for embryo rank and recipient age. Embryos characterised by the AI assistant as being either Optimal or Good had an SA incidence of 16.1% (CI: 10.7 - 23.6%, n=20/124) as opposed to 25.0% for embryos scored as either Fair or Poor (CI: 14.9 - 38.8%, n=12/48, P=0.005, OR=0.46). The incidence of SA decreased with maternal age (21.3% in ≤35, 17.9% in 36-38, 16.4% in ≥39, P=0.0181, OR=0.354). However, this effect could be explained by an increase in the proportion of donor eggs employed in the older patient group, so that recipient age might not necessarily reflect the age of the oocyte provider. A significant interaction between embryo ranking and recipient age (P=0.018) (figure 5.2) was also detected. These findings are presented in Table 5.3.

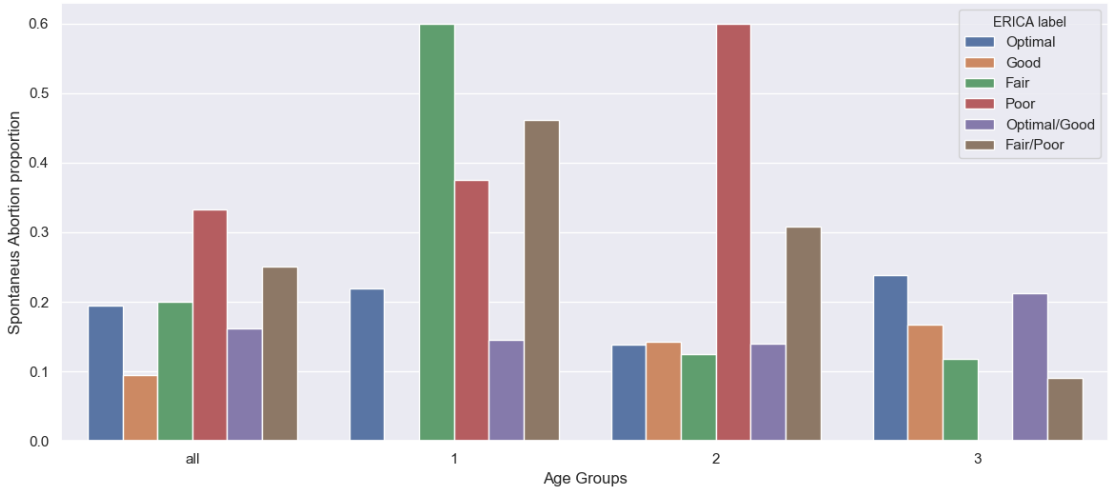
	≤35 years		36-38 years		≥39 years	
Embryo Rank	**n**	**SA**	**n**	**SA**	**n**	**SA**
Optimal or Good	48	7 (14.6%)	43	6 (14.0%)	33	7 (21.2%)
Fair or Poor	13	6 (46.2%)	13	4 (30.8%)	22	2 (9.1%)
Overall	61	13 (21.3%)	56	10 (17.8%)	55	9 (16.4%)

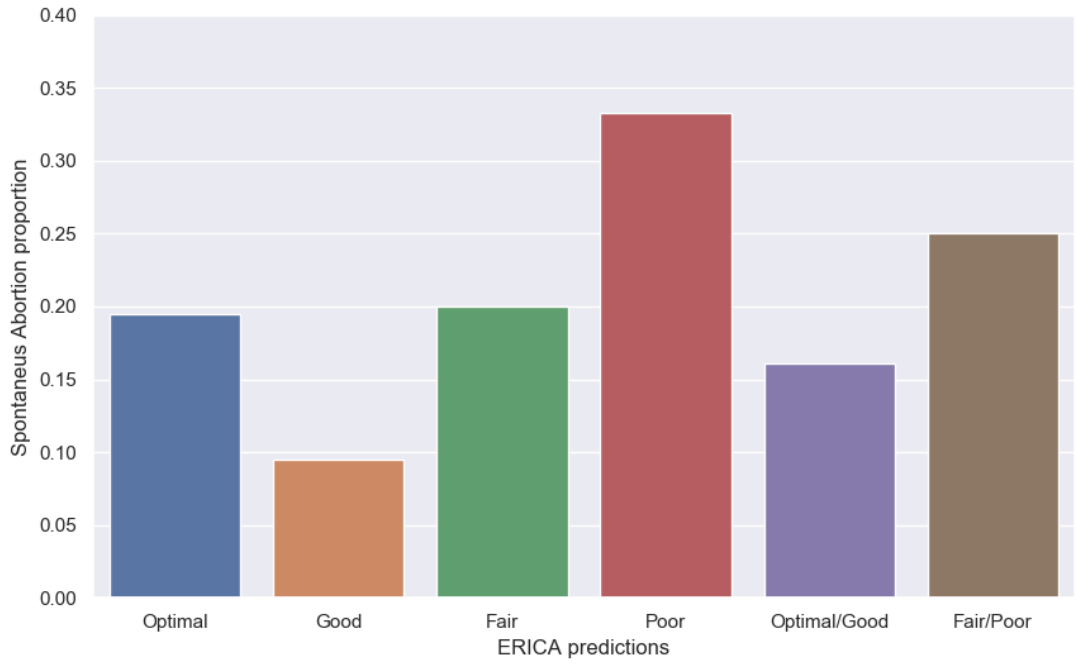
TABLE 5.3. Incidence of SA by ERICA™ embryo rank and recipient age

Overall, the results support a correlation between risk for SA and ERICA™'s embryo ranking. The calculated accuracy of this prediction (the overall probability that a biochemical pregnancy was correctly predicted to progress or not) was 67.4% (CI: 59.9 - 74.4%). The test also achieved good

sensitivity (74.3%, CI: 66.2% - 81.3%) but poor specificity (37.5%, CI: 21.1 - 56.3%), suggesting that the AI was better at assigning an embryo that was associated with an ongoing pregnancy to the Optimal or Good rank (sensitivity) than it was at assigning an embryo that was associated with SA to the Fair or Poor rank (specificity).

FIGURE 5.2. Spontaneous abortion proportions by age groups and ERICA scoring (above, broken down by age group, below overall)





5.4.2. SA in euploid embryos determined by PGT-A

A sub-analysis compared the same embryo ranks and groups for known euploid embryos (n=50). The incidence of SA in euploid embryos ranked as either Optimal or Good was 13.3% (CI: 5.3 –

29.7%, n=4/30), and 20.0% in embryos ranked as either Fair or Poor (CI: 8.1 – 41.6%, n=4/20); the difference between these two proportions was not statistically significant (P=0.531).

5.5. Discussion

SA is a common complication encountered during fertility treatment, with an incidence of 12% to 22% in IVF cycles (Bu 2020, Yang 2019), although PGT-A has been shown to lower the risk of SAs following the transfer of euploid embryos (Rubio 2017, Lee 2019). SAs impose a high toll on patients not only emotionally, financially, and physically but also by extending the time to live birth, which for some patients, such as those with severely compromised ovarian reserves, could further reduce their overall chances of conceiving at all.

Since current ART treatments tend to generate more than one transferrable embryo, the embryo selection process to improve the time to pregnancy (usually using the subjective judgment of the embryologist) has long merited attention (Roque 2020, Ferrick 2019). As a consequence, approaches to improve this process using more objective and accurate ways of selecting viable embryos have been proposed and include PGT-A (Munné 2019), time-lapse evaluation (Kovacs 2014), visual classifications (Gartner 2000), and, more recently, AI (Dimitriadis 2021). Despite the controversial nature of PGT-A in its current form, it is the closest to a gold standard that we have for determining aneuploidy potential and, hence, chances of an ongoing pregnancy past the first trimester.

Embryo quality is largely considered the primary factor responsible for achieving a positive pregnancy test, although several other factors, including endometrial receptivity, may contribute (Ashary 2018, Tıras 2014). Embryos with severe aneuploidy, such as those with monosomies and chaotic embryos, are likely to be lost before a positive pregnancy test is achieved. Beyond this, one could assume that many of the basic factors that might have prevented implantation have been successfully overcome, such as the implantation window and the embryo transfer procedure. At this point, aneuploidy (principally trisomy of the autosomes and monosomy X) is relevant. Although the causes of SA are multifactorial, aneuploidy is widely cited as the most common reason (Hassold 2001, American College of Obstetricians and Gynecologists' Committee on Practice Bulletins 2020).

ERICA™ was developed as a ranking tool to advise embryologists about the best order in which embryos from a cohort could be transferred. This ranking was developed using PGT-A results as its "ground truth," with clinical results supporting its performance (Chavez-Badiola 2020b). Although ERICA™ was not designed to predict the risk of miscarriage per se, aneuploidies are nonetheless the most common cause of SA. So, in this study, we determined whether there was a correlation between ERICA™'s scores and SA. Indeed, we observed that ERICA™'s ranking assessment was somewhat predictive of SA as it was able to classify just over two-thirds of the embryos analysed correctly.

Nonetheless, we acknowledge that before a truly robust statistical evaluation of our hypothesis can be made, the limited sample size achieved in this study should be addressed. Furthermore, while we do not claim that ERICA™ is a more accurate assessment of embryo ploidy than the gold standard PGT-A, it was interesting to note a potential application for this tool in SA prediction. We must also consider that PGT-A usually requires cryo-storage and subsequent frozen embryo transfer, which are expensive and invasive procedures for the embryos. Not all embryos will survive the biopsy and warming process, and some couples only have fewer embryos per cohort.

A better understanding of the likelihood of pregnancy and miscarriage for each embryo will assist clinical staff in managing their patients' expectations during counselling. For future studies with a miscarriage focus, karyotyping of the pregnancy tissue would establish whether ERICA™ is as good as, or nearly as good as, PGT-A for predicting the ploidy status and, hence, the likely SA potential of an IVF cycle.

6. Automated identification of blastocyst regions at different development stages (specific aim d)

The following chapter is encapsulated in a prior published work:

Farias AF, Chavez-Badiola A, Mendizabal-Ruiz G, Valencia-Murillo R, Drakeley A, Cohen J, Cardenas-Esparza E. Automated identification of blastocyst regions at different development stages. Scientific Reports. 2023 Jan 2;13(1):15.

My contributions include the original idea to identify relevant blastocyst structures through different development stages, building databases, literature review, analysis of results and discussion.

Some of the specific wording is changed to give context to a thesis chapter. However, the majority remains unchanged as the words were originally authored by myself.

6.1. Chapter Summary

The selection of the best single blastocyst for transfer is typically based on the assessment of the morphological characteristics of the zona pellucida (ZP), trophectoderm (TE), blastocoel (BC), and inner cell mass (ICM), using subjective and observer−dependent grading protocols. We present the first automatic method for segmenting all morphological structures during the different developmental stages of the blastocyst (i.e., expansion, hatching, and hatched) is herein proposed. The database contains 592 original raw images that were augmented to 2132 for training and 55 for validation. The mean Dice similarity coefficient (DSC) was 0.87 for all pixels, and for the BC, BG (background), ICM, TE, and ZP was 0.85, 0.96, 0.54, 0.63, and 0.71, respectively. Additionally, the method was tested against a public repository of 249 images, resulting in accuracies of 0.96 and 0.93 and DSC of 0.67 and 0.67 for ICM and TE, respectively. A sensitivity analysis demonstrated that this method is robust, especially for the BC, BG, TE, and ZP. It is concluded that this approach can automatically segment blastocysts from different laboratory settings and developmental phases of the blastocysts, all within a single pipeline. This approach could increase the knowledge base for embryo selection.

6.2. Chapter introduction

In-vitro fertilisation (IVF) is one of the most common and effective methods for the treatment of infertility. This procedure consists of stimulating a woman’s ovaries to generate multiple eggs in a given cycle. The mature eggs are retrieved and placed in a Petri dish, fertilised by sperm and cultured in an incubator under controlled environmental conditions (Gartner 2000). In the days to follow, successfully fertilised eggs, now embryos, will undergo different stages of development until after five to six days, some will reach the blastocyst stage. A blastocyst may then be transferred to the woman’s uterus to generate a pregnancy.

The blastocyst is the first morphologically differentiated state of the human pre-implantation embryo, in which cellular structures are arranged in at least four regions: the trophectoderm (TE), which is a layer of cells that surrounds a fluid cavity known as the blastocoel (BC), the embryoblast or inner cell mass (ICM), and the zona pellucida (ZP), which is a protective layer. The development stage of the blastocyst also might be defined by at least three phases: expansion and thinning of the zona pellucida, hatching of the embryo through the zona pellucida, and hatched from the zona pellucida referring to the process of leaving the ZP. These stages are part of embryologists' criteria when

evaluating an embryo's quality (Gartner 2000). Figure 6.1 depicts the structures of expansion, hatching and hatched blastocysts.

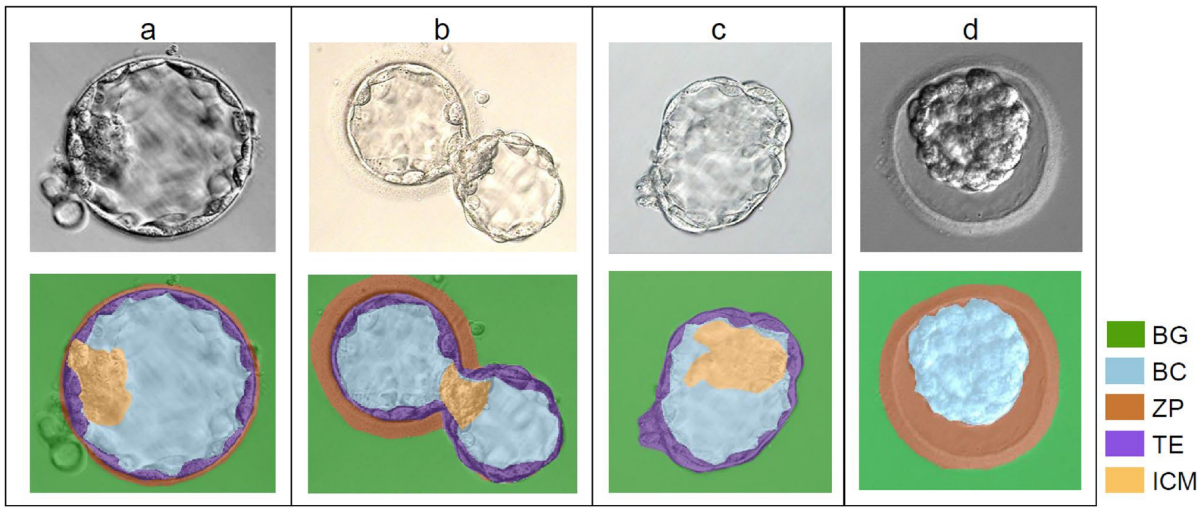


FIGURE 6.1. Examples of the blastocyst structures at each developmental stage: expansion (a), hatching (b), hatched (c) and collapsed (d). Colours indicate the regions of interest in the micrograph: background(BG), blastocoel (BC), zona pellucida (ZP), trophectoderm (TE), and the embryoblast or inner cell mass (ICM).

In many IVF treatments, multiple blastocysts are available for transfer. While transferring multiple blastocysts is sometimes considered as an option to increase the chances of a successful pregnancy, this practice is not recommended due to the possibility of multiple pregnancies and its associated risks, including premature birth, the need for a cesarean section, and a higher risk for pregnancy loss, and other maternal and neonatal morbidities. In that context, a single embryo transfer is a better alternative, necessitating improving embryo classification. The selection of the best blastocyst for transfer is commonly based on the assessment of morphological characteristics and rate of development. However, there are more invasive and expensive approaches, including embryo biopsy and genetic testing, that require cryopreservation. The most common way to perform morphological assessment is by visually inspecting the embryos or using a digital image source. The assignment of quality scores, such as those proposed by Dokras and Gartner, is advantageous (Dokras et al.1993, Gartner 1999). Still, a significant limitation of this approach is subjectivity, which is related to judgement, training, and expertise.

Automatic identification of the discrete regions of a blastocyst using digital microscopy images could aid in overcoming the drawbacks of historical methods by increasing the objectivity and reproducibility of the embryo selection process. These computer-based analysis tools could provide valuable quantitative information to the embryologist to support and improve the decision-making process during an IVF treatment. For instance, based on the intensity patterns of the regions in the

image of a blastocyst, it is possible to automatically infer the embryo quality or its potential by using artificial intelligence (AI) (Chavez-Badiola 2020a; Chavez-Badiola 2020b).

Current work on the automatic, regional segmentation of blastocysts from microscopy images can be divided into two categories. The first category corresponds to methods that rely on computer vision filters and segmentation methods such as watershed segmentation (Saeedi et al.2017, Rocha et al.2017b) or parametric curves such as ellipses (Yee et al. 2013), active contour models (Rad et al.2017,), and level sets (Filho et al.2012, Singh et al.2014). The second type of method is based on the use of supervised machine learning classifiers that are capable of predicting a class label to each pixel of the image according to the pattern contained in a feature vector that is generated using computer vision filters (Kheradmand et al. 2016, Rocha et al. 2017a, Rad et al. 2018a). Most recent methods in this category rely on deep convolutional neural networks (CNN) to automatically determine the best features to extract from the image and perform the segmentation by defining a label for each pixel using encoder-decoder architectures (Kheradmand et al. 2017, Rad et al. 2018b, Harun et al. 2019a, Harun et al. 2019b, Rad et al. 2019a, Rad et al. 2019b, Rad et al. 2020).

Despite the existence of these methods, the automatic segmentation of the blastocyst continues to be a challenging task due to the large variability of shapes and image intensity patterns of each blastocyst region during the different stages of late embryo development, including the expansion, hatching, and hatched stages (See Fig. 1). Furthermore, most of the reports in the literature focus on performing the segmentation of a single region of the blastocyst [e.g., ICM (Rocha et al.2017b, Kheradmand et al.2017, Rad et al.2017, Rad et al.2018) TE (Singh et al.2014, Rad et al.2020), ZP (Yee 2013, Rad 2018a). Only some are designed to perform the simultaneous segmentation of two [e.g., TE and ICM (Saeedi 2017, Harun 2019a)], three [(e.g., ZP, TE, ICM (Filho 2012)] or four (e.g., Kheradmand 2016, Rad 2019a) regions of interest using blastocyst micrographs. Additionally, and to the best of our knowledge, only two works have included hatching embryos. Still, they only segmented the background from the embryo, and only one work (Harun 2019b) included images from different laboratory settings (e.g. microscopes and cameras, but not magnifications).

6.2.1. Chapter aims

The aims of this chapter were as follows:

Specific aim di: To develop a method for the fully automatic simultaneous segmentation of the four blastocyst regions (ZP, TE, BC, and ICM) and the BG (culture medium) from digital microscopy based on the use of computer vision filters and supervised machine learning classifiers, including deep learning methods.

Specific aim dii: To perform a sensitivity analysis to test the hypothesis that this method is robust, especially for the BC, BG, TE, and ZP.

6.3. Material and Methods

As an overview, the materials and methods can be summarised into the database description, the segmentation procedure (images pre-processing, pixel classification, and segmentation refinement), and testing the model (sensitivity analysis and validation through a testing set and a public repository) as described in Fig. 6.2

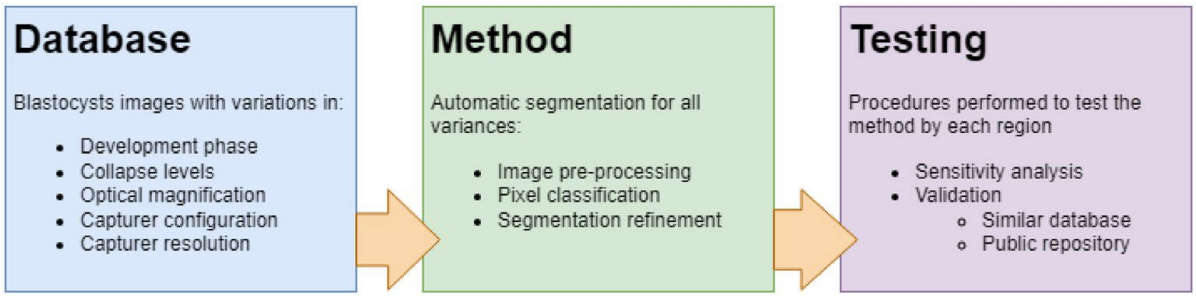


FIGURE 6.2. Overview of the materials and methods used in the study

6.3.1. Database and ground truth description

This work employed blastocyst images generated through data augmentation techniques from a dataset of 592 blastocyst images collected from two IVF clinics for six consecutive months. Informed consent was obtained from all subjects. The image data collection and all experimental protocols were performed in accordance with relevant named guidelines and regulations (IRB approval number RPA-2021-03, IRB name: Comité de Bioética en Investigación New Hope Fertility Center, Registry number: CONBIOETICA 09-CEI-00120170131). The images were collected in two different laboratory settings with the following equipment respectively: two inverted microscopes, models IX73

and IX71 (Olympus, Japan), with the capturers LW1135C and DSP3600 series MOD301213 (Hamilton Thorne, Canada), with the magnification objectives LUCplanFLN 20X/40X and LCACHN20X/40X (Olympus, Japan). Embryos were cultured in Continuous Single Culture Complete with HSA (Irvine Scientific, Fujifilm, USA) culture medium. We defined a hatching embryo as a blastocyst with at least one blastocyst cell outside the ZP and a hatched blastocyst as one with most or all blastocyst cells outside the ZP. The database contained images of 372 expansion, 199 hatching, and 21 hatched blastocysts. This imbalance is because the laboratory that provided the images rarely takes the embryos to the hatched stage but performs embryo transfer or freeze during expansion or hatching phases. The only exclusion criterion is that the focal plane was at the thickest part of the embryo (middle layer) so that the trophectoderm cells could be sharply observed. It was not a requirement that the inner cell mass be clearly observable. It is relevant to note that the clinic performs assisted hatching on most of its embryos (per internal protocol). This procedure induces the hatching process to occur sooner than for an embryo without this procedure. For each image, a senior embryologist defined a bounding box containing the blastocyst with self-developed software and then manually segmented the regions of interest (i.e. BG, BC, ICM, TE, and ZP) where possible, also with self-developed software.

In 107 micrographs, the embryologists reported that they were either unsure about the boundaries of the ICM or that it was not visible at all, and they were instructed not to mark any ICM. Also, most of the hatched embryo images lacked a visible ZP.

The result of these procedures is a bounding box for each blastocyst (the smallest possible rectangle defined by the upper left and lower right coordinates that contains the whole blastocyst) and a set of masks (binary images where ‘1’ means that the pixel belongs to the label of the mask and ‘0’ otherwise) with the segmented regions where possible.

6.3.2. Image pre-processing and model training

The method consisted of three steps: (i) pre-processing of the micrographs to segment and standardise the images, (ii) classification of the pixels in the pre-processed image to the relevant blastocyst zone, and (iii) a post-processing refinement to improve the segmentation based on the structure of the blastocyst regions.

Prior to pre-processing, we used random sampling to divide the database into a training set using 90% of the micrographs and a testing set employing the remaining 10%. Data augmentation techniques were performed on the training set by randomly creating modified versions of each image by variations of brightness (by factors of 0.8, 0.9, 1.0, 1.1 or 1.2), by performing horizontal and vertical flips, and rotations (0, 90, 180, or 270 degrees). After this process, we obtained a training set of 2132 blastocyst images and a testing set of 55 (69% expansion, 24% hatching, and 5% hatched).

6.3.2.1. Image pre-processing

Apart from the embryo of interest, blastocyst micrograph images might contain other cells or the instruments employed to manipulate them. Also, the blastocyst position on the image might not be centralised. Therefore, the first part of the pre-processing step consists of identifying the micrograph image's minimum region containing the blastocyst and excluding extraneous objects.

We employed a Python implementation for this procedure to detect objects using a deep neural network architecture named “Retinanet” (Lin et al.2017b, Lin et al.2017a) (available at: https://github.com/fizyr/keras-retinanet). This network model is trained using a dataset of micrograph images where an expert has manually annotated two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$, defining the best bounding box that contains the blastocyst of interest (Fig. 6.3). We employed a transfer learning approach (Shin et al.2016) to train the retinanet, which initialises the weights of the model with those obtained from the training with the ‘Imagenet’ database (Deng et al.2009). Then, the network training was performed using our blastocyst micrograph database for 50 epochs of 1,000 steps each. These parameters were manually set, ensuring that the epochs *vs* accuracy/loss (learning) graph reached a plateau.

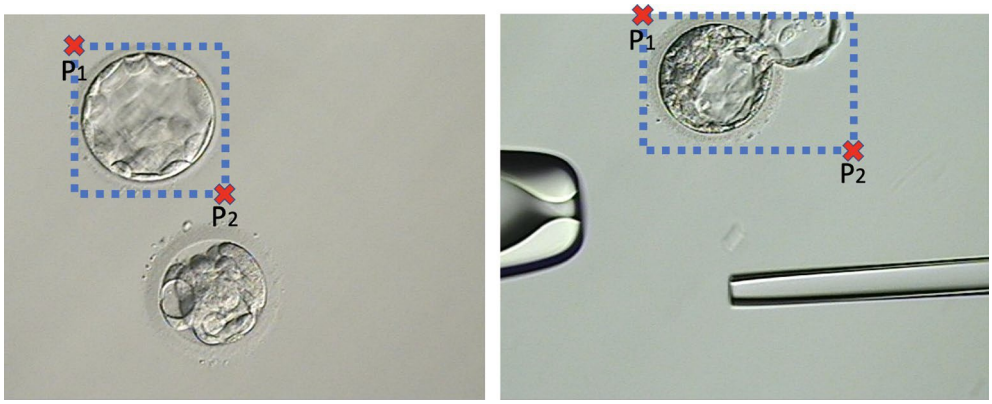


FIGURE 6.3. Examples of two blastocyst micrographs and the bounding box defined by P_1 and P_2.

After training, the network model could automatically define the best bounding box for any input image containing a blastocyst. Using these values, we can automatically crop the original image to obtain an image O with the blastocyst centrally located and contained within the minimum area (smallest possible rectangle).

Blastocyst pictures can be captured using a variety of light conditions or magnifications. Therefore, the second part of the pre-processing step consisted of (a) reducing the variation in brightness of the blastocyst regions and (b) standardising the images to a common scale.

The reduction of variation in brightness was accomplished by applying a gamma correction algorithm to adjust the overall brightness of an image to avoid regions that are too dark or bright (Poynton 2012). Then, the images were normalised using a max-min normalisation, which adjusts the range of the pixel values from 0 to 255. The gamma correction employed consists of four steps: (i) The median intensity value of all the pixels in the images is computed; (ii) If the median value is between 0.45 and 0.75, the gamma transformation is not performed otherwise, steps (iii) and (iv) are performed until the median value is between 0.45 and 0.75, or three gamma transformations are used; (iii) A gamma transformation is carried out on the pixel values using a gamma value computed using the following function $\gamma = 1.25 * \text{median} + 0.375$; and (iv) A new median value is computed from the transformed image and the process is restarted at step (ii).

The standardisation of O was achieved by applying an image scaling transformation using a bilinear interpolation with a scale factor value. Note that it is defined in such a way that each pixel corresponds to one micrometre according to the magnification factor used when the micrograph was taken.

6.3.2.2. Pixel classification

Let I denote the result of pre-processing O. The second step consists of generating five binary images so that each pixel of the images indicates its correspondence to each of the regions of interest: zona pellucida $L_{ZP}(x, y)$, trophectoderm $L_{TE}(x, y)$, blastocoel $L_{BC}(x, y)$, inner cell mass $L_{ICM}(x, y)$, and background $L_{BG}(x, y)$. These binary images are generated by employing a neural network classifier that determines the class of each pixel according to a feature vector that describes the image-intensity patterns in its vicinity.

The feature vector describing the textural characteristics of each pixel is built by applying 21 filter operations (e.g., entropy, Gaussian blur, Laplacian, and Sobel) on the pre-processed blastocyst micrograph. Then, each of the 21 resulting images is convolved with 41 two-dimensional kernels of size 5 × 5, designed to extract the texture patterns of patches of different sizes around the centre pixel (Laws1980). As a result, a feature vector of 861 dimensions is generated for each pixel in the micrograph. Figure 6.4 depicts the process to compute the texture feature vector for each pixel of the pre-processed image I.

For training the network that classifies each pixel according to its vicinity described above, we randomly selected 50 pixels per blastocyst region of interest for each micrograph and generated their feature vectors as described previously. This process resulted in a database of 300,400 data points that were used to train a neural network (three layers with 400 nodes each, Adam optimiser and categorical cross-entropy as the loss function) during 500 epochs with a validation split of 10% of the training data-size, a callback of learning rate reduction (factor of 0.5, patience of 25 epochs, and validation loss as the monitor) and an early stopping callback (patience of 60 epochs, validation loss as monitor, and a minimal delta of 0.001).

6.3.2.3. Segmentation refinement

Depending on the blastocyst region's texture patterns, the neural network could assign an incorrect class to certain pixels. To improve the results, we incorporated a refinement step to account for the topological characteristics of each embryo region. For example, the ZP is the outer layer, which could be visually (but not structurally) connected to the BG and TE regions, the ICM is contained in the BC, and the TE is located between the ZP and the BC or ICM. To add this knowledge to the proposed method, we employed an encoder-decoder scheme using a deep neural network architecture consisting of three convolutional and three max-pooling layers for the encoding and four convolutional and three up-sampling layers for the decoding (See Fig. 6.5). This network is trained using as input a tensor conformed with a concatenation of each of the resulting binary segmentation images S. Similarly, a tensor containing the five-ground truth binary masks G_{BG}, G_{TE}, G_{ZP}, G_{BC}, and G_{ICM} was set as a target. The proposed model was trained using the mean squared error as the loss function, RMSprop as the optimiser, early stopping, and learning rate reduction call-backs during 100 epochs, with a validation split of 10% of the training data size. Mean squared error was used since we observed a better performance than categorical cross entropy (data not shown).

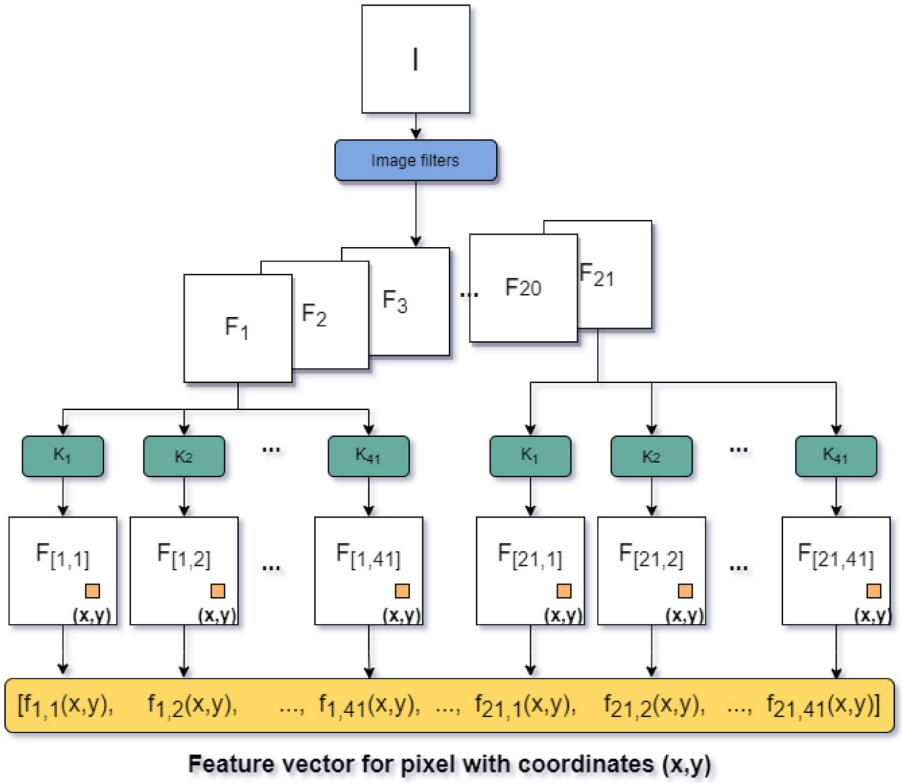


FIGURE 6.4. Depiction of the feature extraction process for each pixel. $[F_1, F2, \ldots, F_{21}]$ are filter operations, the symbol denotes convolution, and $[k_1, k_2, \ldots, k_41]$ correspond to texture extraction kernels.

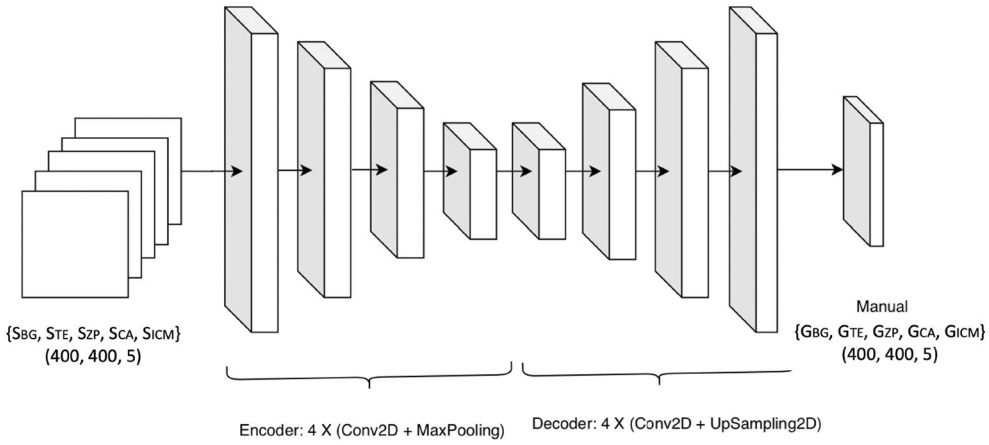


Figure 6.5. Depiction of the encoder-decoder architecture employed to refine the segmentation result.

The validation of the proposed method was measured according to the Dice Similarity Coefficient (DSC) (Dice, 1945) between the ground truth defined by the manual segmentation images and the segmentation results employing the proposed approach. DSC is a standard metric to determine the similarity of two segmentation results. It is computed as two times the area of overlap divided by the total number of pixels in both images. The value of the DSC indicates overlap between regions, ranging from 0 (indicating no spatial overlap) to 1 (indicating complete overlap).

6.3.3. Sensitivity analysis and validation (specific aim dii)

We performed a sensitivity test to evaluate the consistency of the proposed model with respect to changes in location, orientation, and intensity of the input micrographs. For this purpose, we used a completely independent dataset of 159 embryo images (122 hatching, 27 expansion, and 10 hatched). The sensitivity was assessed by creating fourteen modified versions of each image using the following parameters.

- Three corresponding to rotations of 90, 180 and 270 degrees.
- Three corresponding to horizontal (H-), vertical (V-), and combined (HV-) flips.
- Eight corresponding to changes in the brightness in the range of [− 100, 100] with increment steps of 25 units.

During the pre-processing procedure, the transformations were performed after the interpolation step, except for the brightness transformation, which was performed after the cropping. The segmentation of each transformed embryo micrograph was compared against the micrograph with no additional transformations using the DSC. To test the generalisability and compare it against other algorithms, we used the state-of-the-art public repository published by (Saeedi et al. 2017), which was not used at any point during the training process. To pre-process this database of 249 blastocyst micrographs, we assumed that their pixel size was 0.5 micrometres since we did not possess that information. Results were compared to those in (Saeedi et al. 2017) using DSC, sensitivity, specificity, precision, and accuracy of the ICM and TE.

6.4. Results

Table 6.1 lists the DSC results of the proposed segmentation method for each region and stage, including a subset of only images with ICM ground truth (ICM2) in the augmented testing set. Figure 6.6. depicts examples of the blastocyst micrographs after pre-processing, before and after the segmentation refinement, and ground truth segmentation.

Region	Mean (std) DSC before refinement	Mean (std) DSC of the whole dataset	Mean (std) DSC of expansion embryos	DSC of hatching embryos	Mean (std) DSC of hatched embryos
BC	0.72 (0.12)	0.85 (0.06)	0.86 (0.06)	0.84 (0.05)	0.80 (0.11)
BG	0.94 (0.04)	0.96 (0.03)	0.96 (0.02)	0.94 (0.04)	0.95 (0.02)
ICM	0.45 (0.26)	0.54 (0.33)	0.55 (0.32)	0.53 (0.35)	0.44 (0.40)
ICM2*	0.55 (0.18)	0.66 (0.22)	0.66 (0.23)	0.69 (0.21)	0.66 (0.17)
TE	0.59 (0.13)	0.63 (0.13)	0.65 (0.14)	0.58 (0.09)	0.57 (0.08)
ZP	0.66 (0.22)	0.71 (0.24)	0.73 (0.20)	0.75 (0.23)	0.30 (0.51)
All	0.80 (0.06)	0.87 (0.04)	0.88 (0.04)	0.86 (0.05)	0.84 (0.04)

TABLE 6.1. Mean and standard deviation of Dice similarity coefficient for each region before and after the segmentation refinement step. ICM2* - Subset of the embryos filtering those where no ICM was identified by embryologists (20 expansion and 15 hatching). The 'All' row includes the five regions and embryos where no ICM was found.

Note that the segmentation result for all regions improved after the refinement step, with an increment of at least 18% of DSC for the BC and ICM regions. Due to the high homogeneity of texture of the BG and BC regions, the proposed method achieved the best results with a mean DSC above 85%. The ICM segment had the worst DSC, followed by TE and ICM2, with high standard deviations. It can be observed in Table 5.1. that TE was better segmented in expansion embryos than in hatching or hatched embryos. Compared to the public database, our model's performance is shown in Table 6.2. This table reports that our procedure is weaker when DSC and sensitivity are assessed but stronger when assessed by specificity, precision, and accuracy.

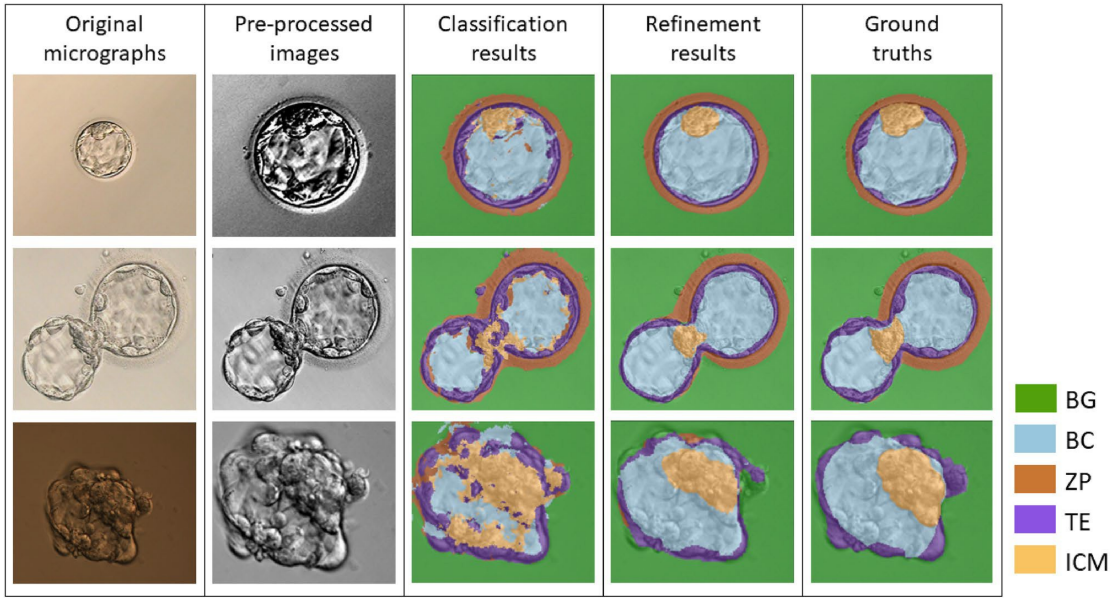


FIGURE 6.6. Examples of the original blastocyst micrographs, the images after pre-processing, before and after the segmentation refinement, and the ground truth. Note these images might have been stretched, shrunk, or cropped for aesthetic purposes, but the pixel values remained unaltered.

Metric	ICM (Saeedi et al.)	TE (Saeedi et al.)	ICM	TE	ZP
DSC	0.79	0.77	0.67	0.67	0.75
Sensitivity	0.84	0.89	0.62	0.59	0.69
Specificity	0.92	0.86	0.99	0.98	0.98
Precision	0.77	0.69	0.87	0.80	0.85
Accuracy	0.91	0.87	0.96	0.93	0.94

TABLE 6.2. Comparison of the results by Saeedi et al. and ours in the same database. *ICM* Inner cell mass, *TE* Trophectoderm, *ZP* Zona pellucida, *DSC* Dice Similarity Coefficient.

Figure 6.7. shows the mean DSC of the regions of interest of the transformed blastocyst images compared with the raw images. Note that the ICM segmentation is the most sensitive to the transformations, followed by the TE region, and BG, followed by BC, is the most robust to transformations. Also, our method is more robust to reduced brightness than increased brightness.

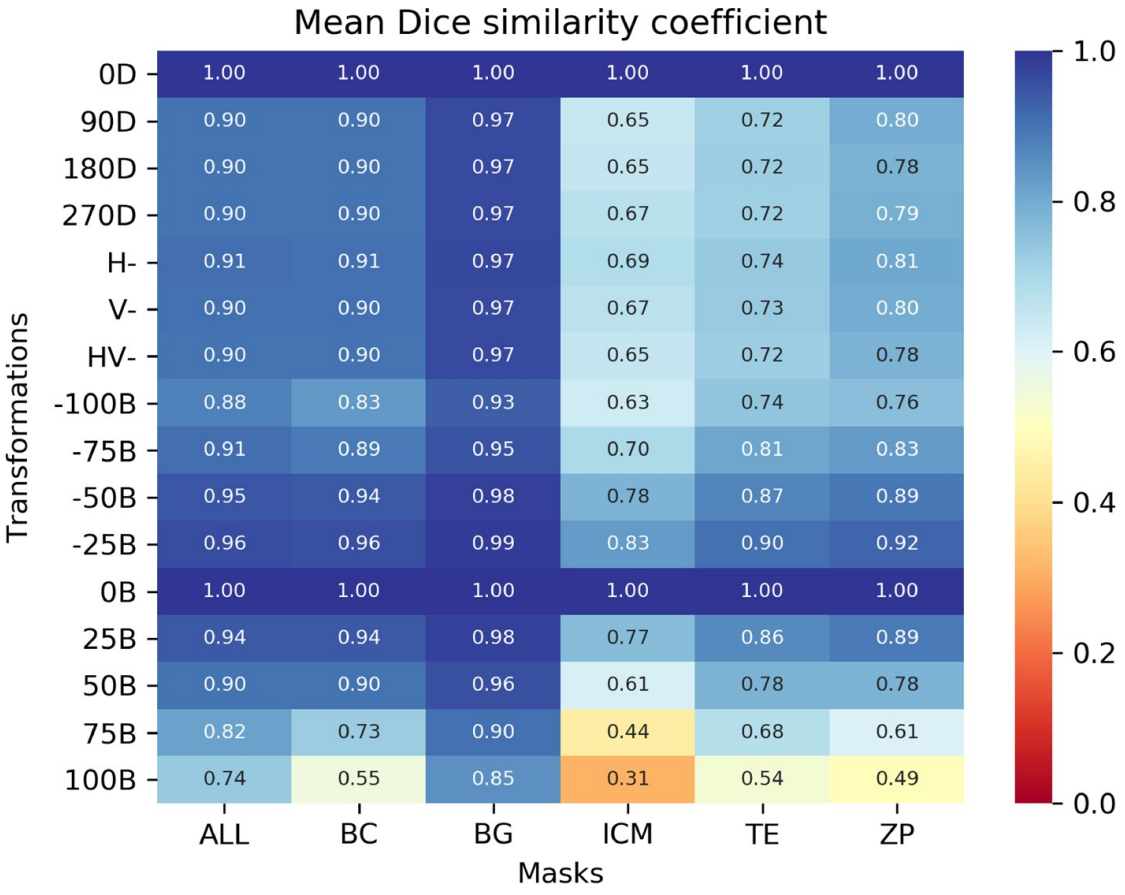


FIGURE 6.7. Mean DSC for the regions of interest (x-axis) between the original blastocyst micrographs and their transformations (y-axis) of the segmentation result. 0D, 90D, 180D, and 270D represents the rotational degrees transformation. H-, V-, and HV- represent horizontal, vertical, or flip transformations. −100B to 100B represent the absolute value of the brightness transformation. *BC* Blastocoel, *BG* Background, *ICM* Inner cell mass, *TE* trophectoderm, *ZP* zona pellucida, *All* all the regions of interest.

6.5. Discussion

This work differs from previously published work in that this method is capable of segmenting blastocyst images from the later developmental stages (expansion, hatching, and hatched) and four different laboratory settings (i.e. cameras, magnifications, light conditions, and general laboratory conditions). The performance of the proposed method was evaluated by computing the Dice similarity coefficient (DSC) between the automatic segmentation results and the manual annotations from a senior embryologist. Its originality resides in the model's robustness to adapt to different laboratory settings to segment blastocysts in any phase (i.e. expanding, hatching, hatched) and in the sensitivity analysis of the performance of the pipeline.

Automatic segmentation of blastocyst regions could potentially overcome drawbacks like inter-/ intra-observer variability and provide novel quantitative tools for blastocyst evaluation, confirming other approaches involving AI-based technologies (Chavez-Badiola 2020a, Chavez-Badiola 2020b). This work presents a novel method for automatically segmenting the relevant biological structures of blastocysts from micrographs taken at different developmental stages: expansion, hatching, and hatched. This method is compared with two ground truth datasets manually segmented by senior embryologists: (i) 592 blastocyst micrographs that included embryos at the three developmental stages with masks for the BG, ZP, TE, BC, and ICM regions from four different laboratory settings (two clinics and two different magnifications each; dataset developed by our group) and (ii) a public repository of 249 expansion embryos and with masks for ZP, TE, and ICM (Saeedi et al.2017). The results showed a mean of 0.87 DSC across all regions of our dataset (with no relevant difference between blastocyst stages) and a mean of 0.79 DSC for the public repository. Additionally, we have performed a sensitivity analysis on an independent database to test the robustness of our method to adapt to different micrographs conditions. We artificially tested different locations and positions of the blastocysts in the micrographs and modified the brightness of the images. However, we understand that there is an inherent limitation and information loss when assessing a 2D image of a 3D structure.

When analysing the performance of this method between different regions, the ZP is highly relevant since it appears in the micrographs as a thin semi-transparent layer with a faint texture that can be difficult to differentiate from the background, especially at its outer limits. Moreover, at the hatching and hatched stages, the ZP could be absent from certain areas of the blastocyst. That might explain why the DSC is near 0.7 since the boundaries of the TE region can be difficult to define. Therefore,

although the qualitative results in Fig. 6.2 appear similar to the ground truth, the area overlap may be less, producing a low DSC. Moreover, in the dataset, we included 6 highly collapsed embryos where no TE was found by embryologists, which made it a harder problem for the AI. The ICM was the most challenging region for embryologists to segment when presented with static images and for our method. This was perhaps because of its large optical similarity with the TE or due to those embryos where it could be erroneously confused with TE either because it was out of focus, completely invisible in the image, or the embryo collapsed. However, the proposed method always tries to find it on the micrographs and produces a zero DSC when absent, which could explain the low mean and high standard deviation. By removing the 107 images with no ICM identified by the embryologists, the score increased from 0.54 to 0.66.

In the second dataset, the method showed higher specificity and precision but lower DSC and sensitivity than the results reported by Saeedi et al. in both regions (TE and ICM). These results suggest a high performance among the pixels predicted as TE and ICM (a high true positive rate) but also leaving many pixels among the mask unselected (high false negative rate). This behaviour is demonstrated by nearly absolute specificity (0.99 and 0.98 for ICM and TE, respectively), interpreted as correctly identifying almost all the true negative pixels, contrasting with a very low sensitivity (0.62 and 0.59 for ICM and TE, respectively). Put simply, when our method classifies a pixel as TE or ICM, it may be true, but there might be many other pixels that our method is missing. However, our method demonstrated a better overall accuracy (metric including both true and false positives) (0.96 and 0.93 for ICM and TE, respectively). When comparing the performance of our method in both datasets, we found a slightly better result in the dataset provided by Saeedi et al., with an increase in DSC of 0.13, 0.04, 0.01, and 0.04 for ICM, ICM2, TE, and ZP. This could be due to our dataset's heterogeneity (laboratory settings, blastocyst phases, embryo quality), and the inclusion of collapsed embryos and non-evident ICM, such as embryos photographed while the ICM was not in the focal plane.

Additionally, further work has been done on the same database. Rad et al., 2018 published a conference paper in which a multi-resolution ensemble of Stacked Dilated U-Net architecture to segment the ICM from expansion embryos, reporting a precision of 88.6%, recall of 91.5%, accuracy of 98.3%, and SDC of 89.5% on their test set (n = 35) (Rad 2018b). A year later, the same group reported BLAST-NET, a Dense Progressive Sub-pixel Upsampling architecture to segment all five segments (Rad 2018b). This architecture showed a Mean SDC of 0.91. Harun et al., in 2019, also reported ICM and TE segmentation performances (one network for each segment) using a Residual

Dilated U-Net. They report that their approach can identify the ICM region with 99.1% accuracy, 94.9% precision, 93.8% recall, and 94.3% Dice Coefficient and the TE region with 98.3% accuracy, 91.8% precision, 93.2% recall, and 92.5% Dice Coefficient (Harun 2019a). Despite the performances of these networks apparently overcoming ours, it is important to highlight that the database contains several limitations; all the images come from a single laboratory setting of expansion blastocysts only. Our dataset contains 4 different lab settings, meaning different magnifications, cameras, and light conditions, and also includes embryos at different phases. Using a two-step architecture, our approach also segments five classes within a single model. Further work would need to be done on different databases and architectures to prove their ability to generalise.

Regarding the sensitivity analysis, the ICM was the most affected by the different image transformations observed, followed by TE, ZP and BC. We can also state that increased brightness was more detrimental to the segmentation than reduced brightness. The rotational and flipping transformations had similar detrimental effects across masks, the ICM being the most affected. The sensitivity analysis highlighted the strengths and weaknesses of our segmentation method, providing a path for improving the robustness of the pre-processing and segmentation in later versions. Further studies could be conducted on larger manually annotated images from different laboratory settings to assess the segmentation model's generalisability properly.

Table 6.1. shows the ability of the autoencoder to refine the segmentation obtained from the texture-based classifier (from 0.80 to 0.87 of DSC for all regions). This novel deep learning architecture shows a hybrid traditional computer vision (top-down) and deep learning approach (bottom-up) of pre-processing and extracting informative features from the images to further classify each pixel for a second step of refinement, including the neighbouring pixels of the segmentation through an encoder-decoder architecture. We foresee that this approach could be applied to other segmentation problems.

Although this method was tested against an external dataset, the generalisability of this model using different protocols, culture methodology, and microscopes remains to be tested. As mentioned above, the blastocysts used for training and validation were developed under a protocol of assisted hatching, which might induce a bias in the method's performance. Embryos treated with this procedure hatch earlier than untreated embryos, altering important morphological aspects of the blastocyst, like the thickness of the ZP, the “figure of eight” shape of the hatching blastocyst, and overall blastocyst size. Other differences in IVF protocols should also be tested, such as culture media, freeze-thawed or

biopsied embryos, and many others. Additionally, a limitation of our results was the poor performance in segmenting the ICM, which might be one of the most relevant features for predictive algorithms due to its high biological relevance. It is also fair to assume a bias of our ground truth dataset, given the proven existence of inter-/intra-embryologist variability in assessing embryos according to their regions (Bormann et al.2020). Finally, the ICM may have to be visualised at different focal positions as the blastocyst is 150-300 micrometres in diameter. Although this is the first work that includes segmenting the three late phases of a blastocyst, there is a strong bias due to the data unbalance. Further work must be done in testing different neural network architectures in a phase-balanced database. The limitations of our work could be summarised as having a moderate performance on segmenting ICM, the high phase imbalance in our database, the bias of the blastocysts images being under an assisted hatching protocol, that the manual annotations were performed by a single senior embryologist and the relatively small database.

While there are methods for blastocyst analysis that do not make use of a segmentation step, the advantage of segmentation is that the data fed to the machine learning models have a structure that depends on the well-identified characteristics of the cell, which can be translated into more transparent and explainable AI models. Future clinical applications using automated segmentation methods might include blastocyst ploidy status prediction, blastocyst transference outcome, or assistance during TE biopsy.

7. Evaluating AI predictive models in reproductive medicine (specific aim e)

The following chapter is encapsulated in a prior published work:

Curchoe CL, Farias AF, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. Fertility and sterility. 2020 Nov 1;114(5):921-6.

My contributions include the general layout, literature review, analysis of results and discussion.

Some of the specific wording is changed to give context to a thesis chapter. However, the majority remains unchanged as the words were originally authored by myself.

7.1 Chapter Summary

Predictive modelling has become a distinct subdiscipline of reproductive medicine, and researchers and clinicians are just learning the skills and expertise to evaluate AI studies. Diagnostic tests and model predictions are subject to evaluation. Their use offers potential for both harm and benefit in terms of diagnosis, treatment, and prognosis. AI models' performance and potential clinical utility hinge on the quality and size of the databases used, the types and distribution of data, and the particular AI method applied. Additionally, when images are involved, the method of capturing, preprocessing, and treatment and accurate labelling of images becomes an important component of AI modelling. Inconsistent image treatment or inaccurate labelling of images can lead to an inconsistent database, resulting in poor AI accuracy. Here, a critical appraisal of AI models in reproductive medicine is discussed, conveying the importance of transparency and standardisation in reporting AI models so that the risk of bias and the potential clinical utility of AI can be assessed.

7.2. Chapter introduction

Every discipline endeavours to improve and optimise its practices to maintain and ensure future competitive advantages, create better products or services, and develop cost-effective production and delivery modes. For this reason, industries regularly pursue the most advanced technological tools being developed by scientists and researchers. As discussed throughout this thesis, AI is an area of computer science that is of intense interest to the health service industries, including reproductive medicine.

As outlined in the general introduction (chapter 1), including AI in reproductive medicine is highly desirable for many reasons. For example, AI systems can make decisions based on facts and supporting data, thus making the decision process reproducible and repeatable; by contrast, humans' decisions may depend on emotional states and are prone to subjectivity, fatigue, and other types of biases (Phillips-Wren G, 2020). Also, AI can learn and analyse very complex patterns with increased resolution while analysing a greater number of variables in comparison with humans (Wang S, 2012). As described in this thesis and elsewhere, AI should reduce healthcare costs while improving outcomes.

Established algorithms can learn quickly from new medical information from successful clinical cases and guidelines to reduce the gap between research and clinical practice. Errors in diagnosis and

treatment in human clinical practice can be reduced by AI assisting in clinical activities and making real-time predictions on health risks. Given full information, AI can identify novel determinants of subfertility, such as specific lifestyle habits, environmental factors, and other key drivers of reduced positive outcomes. If pregnancy and the success of an in vitro fertilisation (IVF) cycle can be predicted, AI can make related healthcare recommendations to increase the chances of success.

While most AI algorithms and computational methods may be difficult to implement, many open-source and free software libraries are now available. However, integrating these models into clinical practice is challenging. Developers should work with a multidisciplinary clinical team to better understand the advantages, costs, requirements, and limitations of implementing such a system; they should also set a risk management plan to identify potentially hazardous scenarios on behalf of patient safety (Tran B, 2019). Potential users of AI models in reproductive medicine include a wide range of individuals: clinicians, embryologists, nurses, and administrative personnel. Newly introduced AI tools must be user-friendly to contribute to an efficient workflow. The potential clinical advantages of such systems should not come at the cost of an increased workload (e.g., increased annotations). An AI model is considered clinically helpful if its decisions can be shown to improve patient outcomes or be economical. ‘‘Overfitting’’—when a model has learned the training data too well and loses its ability to generalise to new data— is a potential risk. When training a model, the goal is to learn an abstraction of the data, not to learn each data point literally. There are frameworks for assessing both the performance of predictive models for decision analysis (net benefit) (Steyerberg E.W, 2010) and for prospective cost-effectiveness analysis (Ulahannan T.J, 2002).

Assuming that an AI has been successfully developed and its clinical recommendations are as good as an experienced senior health care provider, any user will have access to this ‘‘expert opinion’’ anytime, without the human biases that may result in suboptimal performance. Thus, these models will have the advantage of being reproducible, scalable, and tireless, and they could potentially be used anywhere in the world, ideally at a low cost. But to do this, AI models must be ‘‘generalisable.’’ That is, a model’s ability to make sense of data that were not part of its training is referred to as its ability to generalise. Sometimes, noise is introduced into training data to improve a model’s generalisation ability—this allows the model to make sense of noisy data. However, the labels must be accurate (of high quality). This could dramatically change current practices in reproductive health, clinic administration, and decision-making processes internally and in front of the patient (Callahan A, 2017).

7.3. Material and Methods

In this review, we investigate how we, as readers and referees, approach AI and machine learning (ML) literature in reproductive medicine and understand the validation and usefulness of a publication from the clinical context. To this end, we performed a bibliographic search on Web of Science, Scopus, and PubMed. The search keywords were ‘‘artificial intelligence,’’ ‘‘machine learning,’’ ‘‘artificial neural networks,’’ ‘‘convolutional neural networks,’’ ‘‘deep learning,’’ and ‘‘Bayesian’’ combined with ‘‘IVF,’’ ‘‘ICSI, ‘‘ART,’’ ‘‘embryo,’’ ‘‘human embryo,’’ ‘‘semen,’’ ‘‘sperm,’’ ‘‘oocyte,’’ and ‘‘egg.’’

7.3.1. Inclusion Criteria

This review broadly surveys studies that propose, develop, or apply AI techniques to predict or classify reproductive data, such as evaluating embryos, oocytes, or semen quality (Supplemental Table 1, available online; www.repro-AI.org). The review includes published proposals and observational studies (cohort, case, cross-sectional, case reports, and series) for clinical and research applications.

7.3.2. Exclusion Criteria

We excluded studies performed with nonhuman species and those published in languages other than English. Reviews, editorials, and congress abstracts were also excluded.

7.3.3. Study Selection

After removing duplicate publications, four investigators independently assessed the titles and abstracts of all articles. Studies that did not meet the inclusion criteria were excluded. Disagreements regarding inclusion were resolved by consensus or with the involvement of a fifth author.

7.4. Results

This analysis looked at 41 full-length, peer-reviewed publications reporting on AI/ML models for various aspects of reproduction. Each study was classified according to model type (Bayesian, support vector machine, neural network) and analysed for sample type, data origin, accuracy, sensitivity, specificity, and data set size, among other features (Supplemental Table 1). The results of this study are available online and will be dynamically updated at www.repro-ai.org, a neutral

consortium of scientists, physicians, and mathematicians advancing artificial intelligence in reproductive medicine.

7.5. Chapter discussion

Some of the most recent and successful AI methods are based on ML paradigms in which the computer determines the optimal parameters of a high-dimensional model (mostly nonlinear) used to complete the required task (Lecun Y, 2015). These optimal parameters are computed by employing numerical optimisation algorithms that attempt to minimise a defined error function that depends on the model's performance to complete the desired tasks evaluated using the examples provided (Werbos P., 1974).

Therefore, the size and quality of the data used to train an AI model are essential for its success. This is especially important in reproductive medicine, where data are not as easily obtained due to data privacy and a lack of structured electronic medical records (medical history, diagnoses, medications, immunisation dates, allergies, laboratory results, and doctor's notes). To overcome this problem, some groups have participated in multicentric collaborations to increase the size of their data, with excellent results. Low-quality and/or small databases can lead to biased models that might not be generalisable across clinics or reproducible. Because AI relies on the quality of the information used for training, transparency about the quality of data sets is paramount to determining a novel AI model’s clinical potential. Due to the data challenges outlined previously, most AI studies in reproduction involve static two- dimensional image or video analysis instead of data-driven decision-making. Therefore, many good practice points and discussions in this review focus on challenges inherent to image analysis.

A challenge for multicentric collaborations is the critical step of image pre-processing and treatment of images before feature extraction (images are isolated from the background and edges detected) and model training. Clumsy pre-processing and federated subsets of data generated from various geographic locations (with different cropping and contrast, taken on a wide variety of cameras and microscopes) may yield an inconsistent database, leading to AI models with poor accuracy. Similarly, inaccurate feature labelling, such as normal sperm morphology, successful fertilisation, or embryo grading, is also critically important. A worldwide common image/video embryo repository (similar to genome sequence databases), along with automated methods to crop, remove patient information, rotate, and size images for reproduction, is a worthwhile goal.

7.5.1. Four reasons to interpret reproduction AI studies with caution

1. No generalisability assessment. Ideally, an AI-based algorithm should be tested for generalisability across clinics and using a wide spectrum of sample types available in clinical practice. However, some studies presented here describe AI-based algorithms trained using data from a single clinic or with very specific inclusion criteria (e.g., a single brand of incubators, healthy patients, age constraints, only day-5 blastocysts), limiting their clinical applicability in other conditions.
2. Unbalanced data. An AI algorithm is trained to find the best performance with the specific data presented. When the training set is highly unbalanced, the model may be prone to intrinsic bias. Thus, the description of the algorithm should include the actions taken to avoid such bias (Adams N.M, 1999, Drummond C, 2006).
3. Small sample size. By nature, AI algorithms should be trained on large databases because a small sample size can compromise the model's performance and introduce overfitting (Vabalas A, 2019).
4. Limited performance metrics were reported. Algorithms for AI in assisted reproduction are usually classification systems that predict whether a biological sample has a good or poor prognosis. These classifiers can be assessed using different metrics; accuracy and area under the curve are the most common. However, with these metrics alone, a reader cannot perform a more profound performance analysis, for example, if the model is more prone to error in one class or the other. A good practice in this area is to be fully transparent by reporting the confusion matrix or presenting a table with the testing case, prediction, and true results.

An AI model's performance is closely related to the database quality it was trained on. Although seemingly obvious, this is important when a user tries interpreting the model's results. For example, let us say that AI model "A" was trained with embryo images taken on day 3 using a good multicentric database to predict its probability of reaching the blastocyst stage. However, a closer look into the data might show that the entire database included embryos inseminated with intracytoplasmic sperm injection (ICSI) and grown in sequential culture media. Is this model generalisable to images taken on days 3.5 or 4? Or is it accurate for IVF-inseminated oocytes? Is it also applicable when different culture media are used? Standard reporting conventions for IVF and assisted reproductive technology studies should provide not only the number of images but also patient details regarding the IVF/ART cycles.

Another limitation is the concept of ‘‘balanced’’ versus ‘‘unbalanced’’ data. When data are not balanced (i.e., an approximately equal number of positive and negative training examples), the AI will tend to recognise, with a higher probability, data that are in the majority. Or even worse, if 90% of the samples are negative, it can reach 90% accuracy by predicting everything negative. It would only be evident with additional reporting, such as confusion matrix or sensitivity and specificity.

All users should be trained to interpret an AI system's results and be aware of the limitations regarding validation procedures and the conditions of the model. For example, even if the AI was trained on an unbalanced data set, any validation testing should be performed with a balanced data set. To overcome this problem, a proper sensitivity analysis should be performed when evaluating an AI model. However, if a user can access a robust AI system or one trained under conditions similar to those in the clinic where it would be implemented, a quality assurance test should be performed before the model is used in real cases.

As far as I am aware, no prospective studies have been published in which an AI model's performance is compared with standard clinical decisions. Technologies with AI/ML have the potential to highly improve fertility treatments in terms of reducing the time to pregnancy and lowering the associated costs.

7.5.2. Reporting and validating results

Validating an ML algorithm consists of evaluating the model’s performance with respect to a defined ground truth (GT). Most of the time, this GT is defined by experts in the field, and the validation is performed retrospectively. However, in some cases, validating the models in prospective studies is possible. In any case, the most common comparison of ML algorithms is the similarity in the performance with respect to what can be expected from one or more human experts.

Also, because reproductive treatment can vary widely, sensitivity analysis and generalisability tests should be performed before the technology is introduced into daily practice. Furthermore, before an AI model’s clinical application, quality assurance should always be considered whenever a model is introduced into a new setting for the first time.

7.5.3. Ground truth and validation set

As mentioned throughout this thesis, GTs are the known outcomes the AI is trained with. Expected outcomes in reproductive medicine are successful treatments (i.e., positive pregnancy test, presence of a heartbeat, live birth), ploidy status of an embryo or gamete, or any middle step indicating that the probability of a treatment's success is increased (e.g. freeze-thaw survival, normal fecundation, normal cleavage, normal blastocyst). Artificial intelligence systems regularly trained to predict one of these outcomes should be assessed in terms of usefulness in a ''real-life'' treatment protocol. Also, the model should be assessed using an independent validation set. Unfortunately, the most relevant clinical outcomes in reproductive medicine, such as live birth, depend on so many factors (e.g., environmental exposures) that unless big data enters the picture, such ambitious outcome measures might not be as realistic as a GT. Even if the outcome measured is suboptimal or a proxy, such as implantation, ploidy (in the case of embryo assessment), achieving two-pronuclear status, and blastocyst formation (for gamete predictions), they may have to suffice for now. Could it be that the GT, for now, should be a senior embryologist or a vote between several senior embryologists?

7.5.4. Database features

Database conditions (inclusion and exclusion criteria) and training size (after exclusions criteria) should be assessed to determine whether a wide spectrum of possibilities is included in the training and test set. This would help a user identify a model's restrictions or limitations. Artificial intelligence models also have frequently been constructed over unbalanced data (true to false rate different than 1), a potential source of bias in the training process (Lobo J.M, 2008, Saito T, 2015, Mazurowski M.A, 2008).

7.5.5. Reproducibility and repeatability

When a validation study is presented, its methodology should be transparent to ensure reproducible and repeatable results. When evaluating a new AI model, sufficient details should be provided to enable independent validation. Validation data sets should be balanced, even if the AI were trained using an unbalanced data set.

7.5.6. Sensitivity and specificity testing

The output from a model is often a probability, indicating the percentage chance for a positive outcome. Usually, a threshold is set such that any probability above 50% is considered a positive

outcome, and any probability below is considered a negative outcome. Sensitivity and specificity values can be calculated for each threshold. A lower threshold value will yield a higher sensitivity and lower specificity; a higher threshold value will yield a lower sensitivity and higher specificity. The nature of this trade-off can be summarised by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

The AUC corresponds to the probability that a randomly selected positive example has a higher output from the model than a randomly selected negative example. The AUC is a commonly used metric for classification algorithms. It must be understood in terms of the ROC curve: the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) when moving the negative/positive threshold. An important caveat is that this metric cannot be trusted in highly unbalanced data (Adams N.M, 1999, Drummond C, 2006).

The performance of a sensitivity test is highly desirable and should be presented when the development of a new model is reported. Sensitivity tests should include the model's accuracy and performance given different conditions of the test set (e.g., age cohorts, culture conditions). Similarly, the performance of an error analysis, in which the errors obtained during the testing phase are assessed, and emerging patterns should be considered.

7.5.7. Results provided

It is also relevant to note how results are presented. An optimal way to present a model's results is through a confusion matrix, allowing readers to compare true versus false positives or negatives and view any metric related to the table (including accuracy, precision, F1 score, sensitivity, specificity, etc.). For example, is the model more prone to false negatives or positives? And how would these results impact the protocols of your clinic? Is a true outcome more confident than a false outcome? What are the implications of the table for a laboratory's protocols?

7.5.8. Separate data for validation and testing

A common strategy for validating models is by n-fold cross-validation. Suppose 100 different models are validated using the same *n*-fold cross-validation data set. In that case, there is a high probability that one of these models will accidentally perform better than the others. Therefore, a separate test set should be used when reporting the final results. The test set should not have been used to evaluate any other models, or else it will suffer from the same problem as the initial *n*-fold cross-

validation, and it should be a balanced data set. This is another reason that a study needs to be reproducible so it can be verified by others using different data sets.

7.5.9. Test set comparison to user settings

As discussed earlier, the limitations of the training database will be reflected in the limitations of a model's results and usability. If a model were not trained with settings similar to where the user intends to adopt it, further studies would need to be performed to assess its quality and interpretability. The impact of the technology's inclusion into the clinical process and patient outcomes must be analysed. Some studies might not be designed directly for use in clinical practice but rather to test an AI model's performance against other methods.

7.6. Reader and Referee Considerations: A Quick Reference

- Is the purpose of the presented study clearly defined (15): development of a new model (without attempting to validate), the validation of a new model, or a combination of both?
- Is any reasoning presented for the choice of sample size (i.e., by studying similar problems in the literature, use of a domain expert, or statistical method)?
- Was an inflection point reached where providing more training data to the system no longer yields improved accuracy (e.g., data size vs accuracy plot)?
- Were enough data analysed to demonstrate that one model is better than another?
- Is the architecture fully described, explaining why each algorithm was selected to help the reader determine the appropriateness of the tests or the comparisons?
- In the case of neural networks, does the description of architectures include the number of layers employed (to differentiate at a glance those algorithms with deep learning capabilities)?
- Are enough data points used to achieve a desired level of performance (enough data to reasonably capture the relationships between input features and between input features and output features)?
- Did the training set achieve a sufficient estimate of model performance?
- Are both controlled and uncontrolled variables identifiable and described? Are the inclusion and exclusion criteria for the data collection specified (which directly reflect in which situations it can be used within a clinical environment)?
- Is it a balanced data set preferred for training and validation?

- If unbalanced data sets are used for training, is it acknowledged, along with a description of the selected mathematical approach used to compensate for such imbalance?

7.7. How do we critically appraise the efficacy of AI in reproductive medicine?

The average reader or journal referee is likely not sufficiently well-versed in the nuances of AI/ML to differentiate between a good and poor AI study. AI is complex, and the reports routinely introduce mathematical or computer science jargon alien to medical science. Thus, AI-related manuscript referees should include at least one ML data scientist or mathematician to ensure robustness. Although AI/ ML studies rely on sophisticated statistical and mathematical analyses, they are presented within a clinical setting we are familiar with. Certain guidelines have been published to help reproduction clinicians better appraise AI technology (Liu Y, 2019; Jaeschke R, 1994).

The first step in appraising an AI paper in the context of reproductive medicine is to understand what the report is about. In brief, it could present (Liu Y, 2019):

1. The development of a new AI/ML prediction model (not yet validated)
2. The validation of a new or an update of an existing model
3. A combination of both

The first scenario focuses on the model's robustness, database characteristics (size, distribution, balance), and an evaluation of the proposed mathematical approach. In the second scenario, there could be a description of a tool that may be implemented in a clinical setting; most traditional ways to evaluate a new or updated predictive model's clinical usefulness would apply, including the study's replicability. The third scenario is probably the most challenging for the clinician because, besides assessing clinical usefulness, evaluating these studies involves understanding applied mathematics and complex statistics. Approaching such studies as if they were presenting two different outcomes in a single document is perhaps the best way to understand the investigators' intentions.

Several tools and checklists have been developed to assess study design, reporting, validation, and risk of bias for reviews and meta-analyses, prediction models, prognostic models, and clinical trials. However, due to the inherent characteristics of AI/ML, it is essential to note that these tools might not

entirely apply to AI/ML studies. Supplemental Table 2 (available online) compares the uses and limitations of several models.

Several tools are particularly relevant for referees and readers of contemporary AI and reproductive medicine studies. The TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis) initiative developed a set of recommendations for reporting studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. It does not describe how prediction modelling studies should be performed but how studies (regardless of how well or poorly designed) should have common reporting elements. The presence or absence of a reported element is not correlated with the risk of bias in the model.

The PROBAST (Prediction model Risk Of Bias ASsessment Tool) tool assesses the risk of bias and concerns regarding the applicability of diagnostic and prognostic prediction model studies. Evaluators can use PROBAST to assess model development and validation studies, including those updating a prediction model. Additional quality assessment tools may be needed depending on the study design (not intended for predictor finding studies, prediction model impact studies, or clinical trials).

Future AI research should be evaluated using tools in development, such as TRIPOD-AI and tools specific for clinical studies like Cochrane RoB (Supplemental Table 2). The TRIPOD steering group is developing additional reporting guidance for prediction model studies based on AI or ML methods (TRIPOD-AI) for model development, validation, or updating.

Randomised controlled trials (RCTs) are considered as close to a gold standard as possible for clinical validation when assessing ‘‘static’’ interventions. Their benefits are obvious in most areas, although several issues inherent to their construction and aims could challenge their applicability in AI/ML studies. The most pressing concern might be the speed at which AI/ML models can be updated and their inherent learning capabilities, which allow for a model to be tuned to specific practice conditions. This contrasts the complex logistics required for an RCT and the long time it takes from design and recruitment until publication. In other words, a 2-year RCT project could demonstrate that a 2.5-year-old version of an AI/ML model is good by the time the same ML model is already in its third or fourth version/update (Sanson-Fisher R.W, 2007, Nichol A.D, 2010).

Because ML models are designed to learn, assessing a model's robustness (e.g., internal validation, external validation) is crucial when considering its clinical readiness. Before a clinical application is considered, evaluating a model's performance after tuning it for individual practices through a standardised quality assurance (QA) process is paramount and should be considered good clinical practice. In other words, a well-constructed ML system should always be allowed to learn from new data sets, and its clinical validation will depend on high standards of QA protocols (Mahadevaiah G, 2020).

The data from this study are now published online at www.Repro-AI.org.

8. General Discussion

This thesis was successful in fulfilment of its specific aims in that:

- A prototype for predicting pregnancy test results after embryo transfer using machine learning was successfully developed by image feature extraction and analysis (AIR E®). The results accept the hypothesis that this novel approach can predict pregnancy test results (specific aim a), indicating the feasibility of using the system to predict positive implantation tests from a single digital image to prognosis. A potential new and practical approach to embryo classification and selection that may be easily integrated into clinical practice was thus developed.
- An Embryo Ranking Intelligent Classification Algorithm ("ERICA" – a natural successor to AIR E®) was developed. ERICA, an AI clinical assistant, was established to have embryo ploidy and implantation predicting capabilities**.** In other words, the hypothesis that ERICA can predict ploidy status in IVF embryos was confirmed (specific aim b). Following training and validation, ERICA was more successful than random selection and experienced embryologists in correctly identifying and ranking embryos with the highest implantation potential using a static picture as the only source of information. ERICA thus has the potential to assist embryologists and clinicians without the need for time-lapse or invasive embryo biopsy.
- The application of ERICA based on static images was further demonstrated, and the hypothesis that it may have the potential to predict first-trimester pregnancy loss/ spontaneous abortion was accepted through a retrospective pilot study (Specific aim c). Specifically, results support a correlation between the risk of spontaneous abortion and embryo rank as determined by ERICA's classification algorithm. The preliminary study suggests that AI (ERICA™), which was designed as a ranking system to assist with embryo transfer decisions, might also help provide information for couples on the risk of spontaneous abortion.
- The first automatic method for segmenting all the morphological structures during the different blastocyst developmental stages was developed. A sensitivity analysis established that this method is robust (Specific aim d). This approach can automatically segment blastocysts from different laboratory settings and developmental phases of the blastocysts, all within a single pipeline and could increase the knowledge base for embryo selection.

- A systematic evaluation of predictive AI models in reproductive medicine is demonstrated (Specific aim e). A critical appraisal of AI models in reproductive medicine is discussed, conveying the importance of transparency and standardisation in reporting AI models so that the risk of bias and the potential clinical utility of AI can be assessed. Four reasons to interpret reproduction AI studies with caution are given, as is a guide to critically appraise AI's efficacy in reproductive medicine. Finally, a quick reference reader and referee consideration guide is given.

This thesis thus demonstrates that AI can be utilised as a promising tool to resolve many longstanding challenges in reproductive medicine, assist clinicians in decision-making, and achieve the ultimate goal of a healthy, live-born baby. Here, attention is paid mostly to embryo analysis; however, similar principles apply to sperm and egg analysis and the interaction between the two gametes. While this thesis focuses on embryo analysis, my work in these other areas is presented in the appendix. In other areas of reproductive medicine, AI has the potential to assist as well, including endometrial receptivity, uterine function, fertility impact of diseases such as endometriosis and adenomyosis, recurrent implantation failure, and recurrent pregnancy loss (Curchoe C 2021). Barriers to achieving this include health record privacy terms, paper records and variations in electronic medical record systems (Hickman 2020). AI will also play a significant role in the future of IVF, which realistically, will only come about if the artisanal approach of manual handling, basic microscopy, and subjective analysis is replaced. Automation and AI are the most likely combinations of modalities to achieve this (Abdullah 2022)

8.1. Promise and Progress of AI in Embryo Assessment

Artificial intelligence (AI) techniques involving machine learning, computer vision and neural networks have shown increasing promise for automated analysis of embryo images and prediction of viability and outcomes in IVF (Curchoe et al., 2020). As demonstrated in this thesis, proof-of-concept models have proven successful in specific controlled environments. Pregnancy likelihood following embryo transfer could be predicted using computational image feature analysis (Chapter 3). At the same time, the proposed ERICA system ranked embryo quality and ploidy potential, surpassing human experts and random chance when evaluated on static images (Chapter 4,5). Automated segmentation algorithms also reliably identified key morphological structures in embryo micrographs

to aid selection, with robustness to variations in orientation, lighting or development phase (Chapter 6).

Collectively, these results establish feasibility and indicate progress on the possibilities of AI-guided decision support for embryologists. If rigorously developed and clinically validated, such tools could enhance efficiency, precision and consistency in the highly skilled task of distinguishing viable embryos. By reducing qualitative subjectivity and human limitations, AI has the potential to expand global access to IVF services through improved outcomes and productivity.

8.2 Bridging the Gap Between Potential and Practice

However, most current applications are limited to proof-of-concepts in tightly controlled environments using small datasets from individual clinics. The paradigm shift promised by AI in embryo assessment will require bridging substantial gaps before real-world clinical deployment. Model reliability and safety must be demonstrated more conclusively across diverse patient populations and IVF protocols.

Large-scale, heterogeneous and multicentre training data will be essential for generalisable AI tools that maintain accuracy despite variations in laboratory or stimulation conditions (Kragh & Karstoft, 2021). Standardised imaging protocols and consistent pre-processing are also key. Robust validation frameworks must keep pace with the complex considerations of continuously self-improving systems reliant on recursive algorithms (Hickman et al., 2020).

User perspectives must inform development so AI solutions integrate smoothly into existing clinical workflows rather than disrupt them. Understanding end-user receptivity via participatory design can circumvent bottlenecks to adoption (Riegler et al., 2021). Initiatives to facilitate open data sharing and coordinated labelling efforts between global embryology networks could also accelerate progress.

8.3 Responsible Translation Into the IVF Laboratory

The transformation of embryo selection by validated AI assistants must occur responsibly. While such tools can enhance clinical decision-making, human expertise remains invaluable and should be complemented, not replaced. Embryologists provide nuanced situational judgements that algorithms may lack. AI solutions should, therefore, align with and amplify rather than automate or eliminate the embryologist’s role.

Transparency regarding source data, development protocols and performance benchmarks will help earn stakeholder trust. Patient benefit rather than commercial potential should anchor the technology's application. Regulatory structures should incentivise continuous improvement cycles but prevent premature adoption until sufficient evidence of safety and efficacy accrues from multicentre randomised trials.

8.4 The Road Ahead: Cautious Optimism Amid Emerging Possibilities

Realising AI's promise in the high-stakes domain of assisted reproduction warrants cautious optimism tempered by scientific rigour. The foundations laid through the investigations compiled in this thesis provide signals of viability, but the scope for improvement persists. As larger datasets proliferate and collaborative infrastructures mature, rapid gains can be responsibly channelled into cost-effective and globally accessible AI-IVF integration. Concrete progress is evident, but hype must not outpace clinical readiness to balance innovation with prudence in this mission.

8.5. Conclusions

The potential of artificial intelligence to transform embryo assessment as a pillar of assisted reproductive technology is evident. From early machine learning models predicting the likelihood of pregnancy to more advanced neural networks ranking embryo viability or automating morphological analysis, this thesis compiles encouraging proof-of-concept explorations. The rigorous evaluations also outline considerations regarding generalisability, standardisation of imaging inputs and transparency of reporting to earn stakeholder trust. However, a broader motivation also underpins these technical investigations – the need to expand global access to IVF services radically.

The current human-intensive IVF model is reaching unsustainability, with overburdened embryologists as one of the bottlenecks preventing treatment at scale. Extrapolating from declining fertility rates, this restriction could have profound societal impacts. Automating aspects of IVF through rigorously validated AI could thus provide a desperately needed solution by increasing efficiency, precision and consistency. Rather than replacing embryologist expertise, AI-guided decisions could redefine clinical roles from mechanical to cognitive, supported by automated technologies handling

repetitive tasks. If responsibly implemented, AI and advances in robotics and microfluidics could drive a new era of IVF productivity to serve millions more aspiring parents worldwide.

Substantial research gaps remain between demonstrating isolated technical viability to orchestrating whole-workflow automation. But the foundations laid through these preliminary investigations, notably around embryo prioritisation, show AI's potential to assist human limits in tackling a problem of profound scale. Methodical bridging of current limitations via interdisciplinary collaboration could launch a new era in in-vitro fertilisation to uncharted frontiers.

9. References

1. Abdullah, K.A.L., Atazhanova, T., Chavez-Badiola, A., Shivhare, S.B., 2023. Automation in ART: Paving the Way for the Future of Infertility Treatment. Reproductive Sciences 30, 1006–1016.

2. ACOG Practice Bulletin. Screening for Fetal Chromosomal Abnormalities, 2020. Obstetrics & Gynecology 136, e48–e69.

3. Adams NM, Hand DJ. Comparing classifiers when the misallocation costs are uncertain. Pattern Recognit 1999;32:1139–47.

4. Adolfsson E, Andershed AN. Morphology vs morphokinetics: a retrospective comparison of interobserver and intra-observer agreement between embryologists on blastocysts with known implantation outcome. JBRA Assist Reprod 2018.

5. Akinsal, E.C., Haznedar, B., Baydilli, N., Kalinli, A., Ozturk, A., Ekmekçioğlu, O., 2018. Artificial Neural Network for the Prediction of Chromosomal Abnormalities in Azoospermic Males. Urology journal 15, 122–125. https://doi.org/10.22037/uj.v0i0.4029

6. Alikani M. Cryostorage of human gametes and embryos: a reckoning. Reprod Biomed Online. 2018;37:1–3.

7. Alikani M, Fauser BCJM, Anderson R, García-Velasco JA, Johnson M. Response from the Editors: time-lapse systems for ART - meta-analyses and the issue of bias. Reprod Biomed Online. 2018 Mar;36(3):293.

8. Allen, B., Agarwal, S., Kalpathy-Cramer, J., Dreyer, K., 2019. Democratizing AI. Journal of the American College of Radiology 16, 961–963. https://doi.org/10.1016/j.jacr.2019.04.023

9. Al-Inany H, Brotherton J (Cochrane Pregnancy and Childbirth Group), Al Wattar BH, Zou H, Aboulghar M, Mansour R, Proctor M. Intra-cytoplasmic sperm injection (ICSI) versus conventional in-vitro fertilisation (IVF) for couples requiring assisted reproductive technology. Cochrane Database Syst Rev. 2021 Aug 5;8(8):CD012345.

10. Amann RP, Waberski D. Computer-assisted sperm analysis (CASA): capabilities and potential developments. Theriogenology. 2014 Jan 15;81(1):5-17.e3.

11. American College of Obstetricians and Gynecologists' Committee on Practice Bulletins—Obstetrics, Committee on Genetics, Society for Maternal-Fetal Medicine. Screening for Fetal Chromosomal Abnormalities: ACOG Practice Bulletin, Number 226. Obstet Gynecol [Internet] 2020;136(4):e48–69.

12. Antczak M, Van Blerkom J. Temporal and spatial aspects of fragmentation in early human embryos: possible effects on developmental competence and association with the differential elimination of regulatory proteins from polarized domains. Hum Reprod. 1999;14:429–47.

13. Apuzzo MLF, Yoshimatsu J, Mika-Grüttner A, Strehler E, Selenko N, Fischer T, Horvat R, Ploner P, Hackl J, Haas J, Thaler CJ, Strowitzki T. Predicting IVF pregnancy success by DNA methylation signatures of gametes and embryos prior to implantation. Nat Commun. 2021 Jul 26;12(1):4480.

14. Arav A, Natan Y, Kalo D, Komsky-Elbaz A, Roth Z, Levi-Setti PE, et al. A new, simple, automatic vitrification device: preliminary results with murine and bovine oocytes and embryos. J Assist Reprod Genet. 2018;35:1161–8.

15. Armstrong S, Vail A, Mastenbroek S, Jordan V, Farquhar C. Time-lapse in the IVF-lab: how should we assess potential benefit? Hum Reprod. 2015;30:3–8.

16. Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. Cochrane Database Syst Rev 2019. Available at: http://doi.wiley.com/10.1002/14651858.CD011320.pub4.

17. Ashary N, Tiwari A, Modi D. Embryo Implantation: War in Times of Love. Endocrinology [Internet] 2018;159(2):1188–98. Available from: https://academic.oup.com/endo/article/159/2/1188/4792933

18. Badiola AC, Mendizabal G, Cohen J, Flores-Saiffe A, Roberto VM, Drakeley AJ. P–096 Real-time ranking of single spermatozoa using artificial vision analysis of complex motility patterns during ICSI aimed at improving fertilization and blastocyst development. Hum Reprod. 2021; 36.

19. Bahadur, G., Homburg, R., Bosmans, J.E., Huirne, J.A.F., Hinstridge, P., Jayaprakasan, K., Racich, P., Alam, R., Karapanos, I., Illahibuccus, A., Al-Habib, A., Jauniaux, E., 2020. Observational retrospective study of UK national success, risks and costs for 319,105 IVF/ICSI and 30,669 IUI treatment cycles. BMJ Open 10, e034566.

20. Bala, R., Singh, V., Rajender, S., Singh, K., 2021. Environment, Lifestyle, and Female Infertility. Reproductive Sciences 28, 617–638. https://doi.org/10.1007/s43032-020-00279-3

21. Bamford T, Barrie A, Montgomery S, Dhillon-Smith R, Campbell A, Easter C, Coomarasamy A. Morphological and morphokinetic associations with aneuploidy: a systematic review and meta-analysis. Hum Reprod Update. 2022 Aug 25;28(5):656-686.

22. Bamford, T., Smith, R., Young, S., Evans, A., Lockwood, M., Easter, C., Montgomery, S., Barrie, A., Dhillon-Smith, R., Coomarasamy, A., Campbell, A., 2023. A comparison of morphokinetic models and morphological selection for prioritizing euploid embryos: a multicentre cohort study. Human Reproduction. https://doi.org/10.1093/humrep/dead237

23. Barbado A, Corcho O. Interpretable machine learning models for predicting and explaining vehicle fuel consumption anomalies. Engineering Applications of Artificial Intelligence. 2022; 115: 105222.

24. Basile N, Elkhatib I, Meseguer M. A Strength, Weaknesses, Opportunities and Threats analysis on time lapse. Curr Opin Obstet Gynecol. 2019 Jun;31(3):148-155.

25. Ben Rafael Z. Endometrial Receptivity Analysis (ERA) test: an unproven technology. Hum Reprod Open. 2021;2021:hoab10.

26. Bhide P, Srikantharajah A, Lanz D, Dodds J, Collins B, Zamora J, et al. TILT: Time-Lapse Imaging Trial-a pragmatic, multi-centre, three-arm randomised controlled trial to assess the clinical effectiveness and safety of time-lapse imaging in in vitro fertilisation treatment. Trials. 2020;21:600.

27. Borght, M.V., Wyns, C., 2018. Fertility and infertility: Definition and epidemiology. Clinical Biochemistry 62, 2–10. https://doi.org/10.1016/j.clinbiochem.2018.03.012

28. Bori, L., Dominguez, F., Fernandez, E.I., Del Gallego, R., Alegre, L., Hickman, C., Quiñonero, A., Nogueira, M.F.G., Rocha, J.C., Meseguer, M., 2021. An artificial intelligence model based on the proteomic profile of euploid embryos and blastocyst morphology: a preliminary study. Reproductive BioMedicine Online 42, 340–350.

29. Bori L, Paya E, Alegre L, Viloria TA, Remohi JA, Naranjo V, et al. Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential. Fertil Steril 2020;114(6):1232–41.

30. Bormann, C. L., Thirumalaraju, P., Kanakasabapathy, M. K., Kandula, H., Souter, I., Dimitriadis, I., Gupta, R., Pooniwala, R., and Shafiee, H. (2020). Consistency and objectivity of automated embryo assessments using deep neural networks. Fertility and Sterility, 113(4):781–787.e1.

31. Bormann C, Kanakasabapathy M, Thirumalaraju P, Dimitriadis I, Souter I, Hammer K, Shafiee H. O-125 Development of an artificial intelligence embryo witnessing system to accurately track and identify patient specific embryos in a human IVF laboratory. Hum Reprod. 2021;36(Supplement_1).

32. Brownstein, J.S., Rader, B., Astley, C.M., Tian, H., 2023. Advances in Artificial Intelligence for Infectious-Disease Surveillance. New England Journal of Medicine 388, 1597–1607. https://doi.org/10.1056/NEJMra2119215

33. Burkov Andriy. The Hundred-Page Machine Learning Book. 2019, 65-66. Chapter 6: "Neural Networks and Deep Learning.”

34. Bu Z, Hu L, Su Y, Guo Y, Zhai J, Sun Y-P. Factors related to early spontaneous miscarriage during IVF/ICSI treatment: an analysis of 21,485 clinical pregnancies. Reprod Biomed Online 2020;40(2):201–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31883882

35. Cabello-Pinedo S, Abdulla HAN, Seth-Smith ML, Escriba M, Crespo J, Munne S,Horcajadas JA, A novel non-invasive metabolomics approach to screen embryos for aneuploidy. Fertil Steril. 2020;114(3):Supplement, E5–E6.

36. Cabitza F, Campagner A, Sconfienza LM. Studying human-AI collaboration protocols: the case of the Kasparov's law in radio- logical double reading. Health Inf Sci Syst. 2021;9:8.

37. Callahan A, Shah NH. Machine learning in healthcare. In: Sheikh A, Cresswell KM, Wright A, Bates DW, editors. Key advances in clinical informatics: transforming health care through health information technology. Cambridge, MA: Academic Press; 2017:279–91.

38. Campbell A, Nayot D, Krivoi A, Barrie A, Jordan K, Jenner L, et al. Independent assessment of an artificial intelligence-based image analysis tool to predict fertilisation and blastocyst utilisation potential of oocytes, and comparison with ten expert embryologists. Hum Fertil. 2021;24(1):46–69

39. Cao P, Zhu Z, Wang Z, Zhu Y, Niu Q. Applications of graph convolutional networks in computer vision. Neural Comput & Applic 34, 13387–13405 (2022).

40. Carson, S.A., Kallen, A.N., 2021. Diagnosis and Management of Infertility. JAMA 326, 65. https://doi.org/10.1001/jama.2021.4788

41. CDC. 2018 Assisted reproductive technology fertility clinic success rates report. In: Prevention CfDCa, editor. 2020. https://www.cdc.gov/art/pdf/2018-report/ART-2018-Clinic- Report-Full.pdf.

42. Chai, J., Zeng, H., Li, A., Ngai, E.W.T., 2021. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications 6, 100134.

43. Chavez-Badiola, A., Acuña, R., 2017. Assessing ooplasm maturity. Reproductive BioMedicine Online 34, 282.

44. Chavez-Badiola, A., Flores-Saiffe Farias, A., Mendizabal-Ruiz, G. et al. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. Sci Rep 10, 4394 (2020a).

45. Chavez-Badiola A, Flores-Saiffe-Farias A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reprod Biomed Online. 2020;41:585–93. (2020b).

46. Chavez-Badiola, A., Mendizabal-Ruiz, G., Farias, A.F.-S., Garcia-Sanchez, R., Drakeley, A.J., 2020. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Human Reproduction 35, 482–482. (2020c)

47. Chavez A, Badiola, Flores-Saiffe A, Valencia-Murillo R, Mendizabal-Ruiz G, Santibañez-Morales A, Drakeley A, Cohen J. P–243 Improving ERICA's (Embryo Ranking Intelligent Classification Assistant) performance. Should we train an AI to remain static or dynamic, adapting to specific conditions? Hum Reprod. 2021;36(Supplement_1).

48. Chavez-Badiola A, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Griffin D, Valencia-Murillo R, Reyes-Gonzalez D, Drakeley AJ, Cohen J. O-235 ERICA (Embryo Ranking Intelligent Classification Assistant) AI predicts miscarriage in poorly ranked embryos from one static, non-invasive embryo image assessment. Hum Reprod. 2021;36(Supplement_1).

49. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5:493–7.

50. Chen, L., Fu, Y., Wei, K., Zheng, D., Heide, F., 2023. Instance Segmentation in the Dark. International Journal of Computer Vision 131, 2198–2218.

51. Cimadomo D, Fabozzi G, Vaiarelli A, Ubaldi N, Ubaldi FM, Rienzi L. Impact of Maternal Age on Oocyte and Embryo Competence. Front Endocrinol (Lausanne) 2018;9:327.

52. Cocianu, C.-L., Uscatu, C.R., Stan, A.D., 2023. Evolutionary Image Registration: A Review. Sensors (Basel, Switzerland) 23.

53. Collins, G.S., Dhiman, P., Navarro, C.L.A., Ma, J., Hooft, L., Reitsma, J.B., Logullo, P., Beam, A.L., Peng, L., Calster, B.V., Smeden, M. van, Riley, R.D., Moons, K.G., 2021. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 11, e048008.

54. Cornelisse, S., Zagers, M., Kostova, E., Fleischer, K., Wely, M. van, Mastenbroek, S., 2020. Preimplantation genetic testing for aneuploidies (abnormal number of chromosomes) in in vitro fertilisation. Cochrane Database of Systematic Reviews. https://doi.org/10.1002/14651858.CD005291.pub3

55. Costa-Borges N, Giralt G, Albó E, Alvarez A, Ramos J, Hernandez I, Luis M, Calderón G, Munne S. O-122 ICSI in a box: development of a successful automated sperm injection robot with external supervision and minimal manual intervention. Hum Reprod. 2021;36(Supplement_1).

56. Cruz M, Garrido N, Herrero J, Perez-Cano I, Munoz M, Meseg- uer M. Timing of cell division in human cleavage-stage embryos is linked with blastocyst formation and quality. Reprod Biomed Online. 2012;25:371–81.

57. Cummins, J.M., Breen, T.M., Harrison, K.L., Shaw, J.M., Wilson, L.M., Hennessey, J.F., 1986. A formula for scoring human embryo growth rates in in vitro fertilization: Its value in predicting pregnancy and in comparison with visual estimates of embryo quality. Journal of In Vitro Fertilization and Embryo Transfer 3, 284–295. https://doi.org/10.1007/BF01133388

58. Curchoe CL, Bormann CL. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. J Assist Reprod Genet. 2019;36:591–600.

59. Curchoe CL, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. Fertil Steril. 2020;114:921–6.

60. Curchoe CL, Malmsten J, Bormann C, Shafiee H, Flores-Saiffe Farias A, Mendizabal G, et al. Predictive modelling in reproductive medicine: where will the future of artificial intelligence research take us? Fertil Steril. 2020;114:934–40.

61. Cummins J, Breen T, Harrison K, Shaw JM, Wilson LM, Hennessey JF. A formula for scoring human embryo growth rates in vitro fertilization: its value in predicting pregnancy and in comparison with visual estimates of embryo quality. J In Vitro Fert Embryo Transf. 1986;3:284–95.

62. Daar, J., Benward, J., Collins, L., Davis, J., Davis, O., Francis, L., Gates, E., Ginsburg, E., Gitlin, S., Klipstein, S., McCullough, L., Paulson, R., Reindollar, R., Ryan, G., Sauer, M., Tipton, S., Westphal, L., Zweifel, J., 2018. Use of preimplantation genetic testing for monogenic defects (PGT-M) for adult-onset conditions: an Ethics Committee opinion. Fertility and Sterility 109, 989–992. https://doi.org/10.1016/j.fertnstert.2018.04.003

63. Dal Canto M, Moutier C, Brambillasca F, Guglielmo MC, Barto- lacci A, Fadini R, et al. The first report of pregnancies following blastocyst automated vitrification in Europe. J Gynecol Obstet Human Reprod. 2019;48:537–40.

64. Dale B, Menezo Y, Coppola G. Trends, fads and ART! J Assist Reprod Genet. 2015;32:489–93.

65. D'Amour A, Heller KA, Moldovan DI, Adlam B, Alipanahi B, Beutel A, Chen C, Deaton J, Eisenstein J, Hoffman MD, Hormozdiari F, Houlsby N, Hou S, Jerfel G, Karthikesalingam A, Lucic M, Ma Y, McLean CY, Mincu D, Mitani A, Montanari A, Nado Z, Natarajan V, Nielson C, Osborne TF, Raman R, Ramasamy K, Sayres R, Schrouff J, Seneviratne MG, Sequeira S, Suresh H, Veitch V, Vladymyrov M, Wang X, Webster K, Yad- lowsky S, Yun T, Zhai X, Sculley D. Underspecification presents challenges for credibility in modern machine learning. ArXiv. 2020; abs/2011.03395.

66. Dash, R.K., Nguyen, T.N., Cengiz, K. et al. Fine-tuned support vector regression model for stock predictions. Neural Comput & Applic 35, 23295–23309 (2023)

67. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee.

68. Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3):297–302

69. Diaz-Gimeno P, Horcajadas JA, Martinez-Conejero JA, Esteban FJ, Alama P, Pellicer A, et al. A genomic diagnostic tool for human endometrial receptivity based on the transcriptomic signature. Fertil Steril. 2011;95(50–60):e1-15.

70. Diaz-Gimeno P, Ruiz-Alonso M, Blesa D, Bosch N, Martinez- Conejero JA, Alama P, et al. The accuracy and reproducibility of the endometrial receptivity array is superior to histology as a diagnostic method for endometrial receptivity. Fertil Steril. 2013;99:508–17.

71. Diaz-Gimeno P, Ruiz-Alonso M, Blesa D, Simon C. Transcriptomics of the human endometrium. Int J Dev Biol. 2014;58:127–37.

72. Dimitriadis I, Zaninovic N, Badiola AC, Bormann CL. Artificial Intelligence in the embryology laboratory: A review. Reprod Biomed Online [Internet] 2021; Available from: https://linkinghub.elsevier.com/retrieve/pii/S1472648321005575

73. Dokras, A., Sargent, I., and Barlow, D. (1993). Fertilization and early embryology: Human blastocyst grading: an indicator of developmental potential? Human Reproduction, 8(12):2119–2127.

74. Domar AD, Rooney K, Hacker MR, Sakkas D, Dodge LE. Burden of care is the primary reason why insured women terminate in vitro fertilization treatment. Fertil Steril [Internet] 2018;109(6):1121–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29935647

75. Dong K, Romanov I, McLellan C, Esen AF. Recent text-based research and applications in railways: A critical review and future trends. Engineering Applications of Artificial Intelligence. 2022;116, 105435.

76. Doody, K.J., 2021. Infertility Treatment Now and in the Future. Obstetrics and Gynecology Clinics of North America 48, 801–812. https://doi.org/10.1016/j.ogc.2021.07.005

77. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is worth 16x16 words: transformers for image recognition at scale. International Conference on Learning Representations 2021.

78. Drummond C, Holte RC. Cost curves: an improved method for visualizing classifier performance. Mach Learn 2006;65:95–130.

79. Edwards R, Fishel S, Cohen J. Factors influencing the success of in vitro fertilization for alleviating human infertility. J In Vitro Fert Embryo Transf. 1984;1:3–23.

80. Enciso M, Carrascosa JP, Sarasa J, Martinez-Ortiz PA, Munne S, Horcajadas JA, et al. Development of a new comprehensive and reliable endometrial receptivity map (ER Map/ER Grade) based on RT-qPCR gene expression analysis. Hum Reprod. 2018;33:220–8.

81. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J, Socher R. Deep learning-enabled medical computer vision. npj Digit. Med. 4, 5

82. Evans GE, Martinez-Conejero JA, Phillipson GT, Simon C, McNoe LA, Sykes PH, et al. Gene and protein expression signature of endometrial glandular and stromal compartments during the window of implantation. Fertil Steril. 2012;97:1365-73 e1 2.

83. Farias, A.F.-S., Chavez-Badiola, A., Mendizabal-Ruiz, G., Valencia-Murillo, R., Drakeley, A., Cohen, J., Cardenas-Esparza, E., 2023. Automated identification of blastocyst regions at different development stages. Scientific Reports 13, 15. https://doi.org/10.1038/s41598-022-26386-6

84. Ferrick L, Lee YSL, Gartner DK. Reducing time to pregnancy and facilitating the birth of healthy children through functional analysis of embryo physiology†. Biol Reprod [Internet] 2019;101(6):1124–39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30649216

85. Fida B, Cutolo F, di Franco G, Ferrari M, Ferrari V. Augmented reality in open surgery. Updat Surg. 2018;70:389–400.

86. Filho, E. S., Noble, J., Poli, M., Griffiths, T., Emerson, G., and Wells, D. (2012). A method for semi-automatic grading of human blastocyst microscope images. Human reproduction, 27(9):2641–2648.

87. Filho, P.P.R., Rebouças, E. de S., Marinho, L.B., Sarmento, R.M., Tavares, J.M.R.S., Albuquerque, V.H.C. de, 2017. Analysis of human tissue densities: A new approach to extract features from medical images. Pattern Recognition Letters 94, 211–218. https://doi.org/10.1016/j.patrec.2017.02.005

88. Finn A, Scott L, O'Leary T, Davies D, Hill J. Sequential embryo scoring as a predictor of aneuploidy in poor-prognosis patients. Reprod Biomed Online. 2010;21:381–90.

89. Fishel S, Campbell A, Montgomery S, Smith R, Nice L, Duffy S, Jenner L, Berrisford K, Kellam L, Smith R, Foad F, Beccles A, 2018. Time-lapse imaging algorithms rank human preimplantation embryos according to the probability of live birth. Reproductive BioMedicine Online 37(3): 304-313.

90. Fishel, S., Campbell, A., Foad, F., Davies, L., Best, L., Davis, N., Smith, R., Duffy, S., Wheat, S., Montgomery, S., Wachter, A., Beccles, A., 2020. Evolution of embryo selection for IVF from subjective morphology assessment to objective time-lapse algorithms improves chance of live birth. Reproductive BioMedicine Online 40, 61–70.

91. Forzano, F., Antonova, O., Clarke, A., Wert, G. de, Hentze, S., Jamshidi, Y., Moreau, Y., Perola, M., Prokopenko, I., Read, A., Reymond, A., Stefansdottir, V., El, C. van, Genuardi, M., Peterlin, B., Oliveira, C., Writzl, K., Houge, G.D., Cordier, C., Howard, H., Macek, M., Melegh, B., Mendes, A., Radojkovic, D., Rial-Sebbag, E., Ulph, F., Jamshidi, Y., 2022. The use of polygenic risk scores in pre-implantation genetic testing: an unproven, unethical practice. European Journal of Human Genetics 30, 493–495.

92. Gandhi S, Mosleh W, Shen J, Chow CM. Automation, machine learning, and artificial intelligence in echocardiography: a brave new world. Echocardiography. 2018;35:1402–18.

93. Gartner DK, Schoolcraft WB. In vitro culture of human blastocysts. In: Jansen R, Mortimer D (eds). Toward Reproductive Certainty: Fertility and Genetics Beyond. Carnforth, UK: Parthenon Publishing, 1999, 378– 388

94. Gartner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. Fertil Steril [Internet] 2000;73(6):1155–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0015028200005185

95. Gandomkar Z, Brennan PC, Mello-Thoms C, Jun 2018. Mudern: Multi-category classification of breast histopathological image using deep residual networks. Artificial intelligence in medicine 88, 14–24.

96. Gandhi S, Mosleh W, Shen J, Chow CM. Automation, machine learning, and artificial intelligence in echocardiography: a brave new world. Echocardiography. 2018;35:1402–18.

97. García-Pascual, C.M., Navarro-Sánchez, L., Ichikawa-Ceschin, I., Bakalova, D., Martínez-Merino, L., Simón, C., Rubio, C., 2023. Cell-free deoxyribonucleic acid analysis in preimplantation genetic testing. F&S Science 4, 7–16.

98. Gartner, D. K., Lane, M., Schoolcraft, W.B., 2000. Culture and transfer of viable blastocysts: a feasible proposition for human IVF. Human reproduction (Oxford, England) 15 Suppl 6, 9–23.

99. Gartner, D. K., Lane, M., Stevens, J., Schlenker, T., Schoolcraft, W.B., 2000. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. Fertility and sterility 73, 1155–1158.

100. Gartner, D.K., Schoolcraft, W.B., 1999. Culture and transfer of human blastocysts. Current Opinion in Obstetrics and Gynaecology 11, 307–311.

101.Glatstein, I., Chavez-Badiola, A., Curchoe, C.L., 2023. New frontiers in embryo selection. Journal of Assisted Reproduction and Genetics 40, 223–234. https://doi.org/10.1007/s10815-022-02708-5

102.Gilligan MA, Creedon DJ, Mahony H, McCarthy G. Time-lapse imaging helps identify embryos capable of developing to the blastocyst stage: a retrospective cohort study. Reprod Biomed Online. 2016;33:131–9.

103.Gliozheni, O., Hambartsoumian, E., Strohmer, H., Petrovskaya, E., Tishkevich, O., Neubourg, D.D., Bogaerts, K., Balic, D., Antonova, I., Cvetkova, E., Rezabek, K., Kirk, J., Sõritsa, D., Gissler, M., Pelkonen, S., Mansouri, I., Mouzon, J. de, Tandler-Schneider, A., Kimmel, M., Vrachnis, N., Urbancsek, J., Kosztolanyi, G., Bjorgvinsson, H., Wingfield, M., Leyden, J., Scaravelli, G., Luca, R. de, Lokshin, V., Karibayeva, S., Agloniete, V., Bausyte, R., Masliukaite, I., Schilling, C., Calleja-Agius, J., Moshin, V., Simic, T.M., Vukicevic, D., Smeenk, J.M.J., Petanovski, Z., Romundstad, L.B., Janicka, A., Calhaz-Jorge, C., Guimaraes, J.M.M., Silva, P.D. e, Korsak, V., Vidakovic, S., Marsik, L., Kovacic, B., Saiz, I.C., Mondéjar, F.P., Bergh, C., Toitot, S., Schneider, M., Isikoglu, M., Balaban, B., Gryshchenko, M., Bridges, E., Ewans, A., Smeenk, J., Wyns, C., Geyter, C.D., Kupka, M., Bergh, C., Saiz, I.C., Neubourg, D.D., Rezabek, K., Tandler-Schneider, A., Rugescu, I., Goossens, V., 2023. ART in Europe, 2019: results generated from European registries by ESHRE. Human Reproduction 38, 2321–2338.

104.Goodman LR, Goldberg J, Falcone T, Austin C, Desai N, 2016. Does the addition of time-lapse morphokinetics in the selection of embryos for transfer improve pregnancy rates? a randomized controlled trial. Fertility and Sterility 105 (2), 275 – 285e.10.

105.Goodson SG, White S, Stevans AM, Bhat S, Kao CY, Jaworski S, et al. CASAnova: a multiclass support vector machine model for the classification of human sperm motility patterns. Biol Reprod. 2017;97:698–708.

106.Gomez E, Ruiz-Alonso M, Miravet J, Simon C. Human endometrial transcriptomics: implications for embryonic implantation. Cold Spring Harb Perspect Med. 2015;5:a022996.

107.Grace K, Salvatier J, Dafoe A, Zhang C, Evans O. When will AI exceed human performance? Evidence from AI experts. J Artif Intell Res. 2018;62:729–54.

108.Green KA, Patounakis G, DeCherney A, Graham J, Tucker MJ, Widra EA, Levy M, Hill MJ. Day 3 embryo transfer (ET) versus pushing to day 5 in patients with few embryos. Fertil Steril 2016;106:e165.

109.Griffin DK, Fishel S, Gordon T, Yaron Y, Grifo J, Hourvitz A, Rechitsky S, Elson J, Blazek J, Fiorentino F, Treff N, Munne S, Leong M, Schmutzler A, Vereczkey A, Ghobara T, Nánássy L, Large M, Hamamah S, Anderson R, Gianaroli L, Wells D. Continuing to deliver: the evidence base for pre-implantation genetic screening. BMJ. 2017 Feb 14;356:j752.

110.Griffin, D.K., Ogur, C., 2018. Chromosomal analysis in IVF: just how useful is it? Reproduction 156, F29–F50.

111.Gruhn JR, Zielinska AP, Shukla V, Blanshard R, Capalbo A, Cimadomo D, et al. Chromosome errors in human eggs shape natural fertility over reproductive life span. Science [Internet] 2019;365(6460):1466–9.

112.Guo, X., Yu, Q., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., Haake, A. R., 2014. From spoken narratives to domain knowledge: Mining linguistic data for medical image understanding. Artificial Intelligence in Medicine 62 (2), 79–90.

113.Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: A review. Neurocomputing. 2016; 187: 27-48

114.Gupta S, Fauzdar A, Singh VJ, Srivastava A, Sharma K, Singh S. A Preliminary experience of integration of an electronic witness system, its validation, efficacy on lab PERFORMANCE, and staff satisfaction assessment in a busy Indian in vitro fertilization Laboratory. J Hum Reprod Sci. 2020;13:333–9.

115.Hajirasouliha, I., Elemento, O., 2020. Precision medicine and artificial intelligence: overview and relevance to reproductive medicine. Fertility and Sterility 114, 908–913. https://doi.org/10.1016/j.fertnstert.2020.09.156

116.Han C, Zhang Q, Ma R, Xie L, Qiu T, Wang L, et al. Integration of single oocyte trapping, in vitro fertilization and embryo culture in a microwell-structured microfluidic device. Lab Chip. 2010;10:2848–54.

117.Handyside, A.H., Harton, G.L., Mariani, B., Thornhill, A.R., Affara, N., Shaw, M.-A., Griffin, D.K., 2010. Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes. Journal of Medical Genetics 47, 651–658.

118.Handyside, A.H., Kontogianni, E.H., Hardy, K., Winston, R.M.L., 1990. Pregnancies from biopsied human preimplantation embryos sexed by Y-specific DNA amplification. Nature 344, 768–770.

119.Handyside, A.H., Lesko, J.G., Tarín, J.J., Winston, R.M.L., Hughes, M.R., 1992. Birth of a Normal Girl after in Vitro Fertilization and Preimplantation Diagnostic Testing for Cystic Fibrosis. New England Journal of Medicine 327, 905–909.

120.Handyside, A.H., Penketh, R.J.A., Winston, R.M.L., Pattinson, J.K., Delhanty, J.D.A., Tuddenham, E.G.D., 1989. BIOPSY OF HUMAN PREIMPLANTATION EMBRYOS AND SEXING BY DNA AMPLIFICATION. The Lancet 333, 347–349.

121.Harun et al.2019a. Harun, M. Y., Huang, T., and Ohta, A. T. (2019a). Inner cell mass and trophectoderm segmentation in human blastocyst images using deep neural network. In 2019 IEEE 13th International Conference on Nano/Molecular Medicine & Engineering (NANOMED), pages 214–219. IEEE.

122.Harun et al.2019b. Harun, M. Y., Rahman, M. A., Mellinger, J., Chang, W., Huang, T., Walker, B., Hori, K., and Ohta, A. T. (2019b). Image segmentation of zona-ablated human blastocysts. In 2019 IEEE 13th International Conference on Nano/Molecular Medicine & Engineering (NANOMED), pages 208–213. IEEE.

123.Hashimoto T, Koizumi M, Doshida M, Toya M, Sagara E, Oka N, et al. Efficacy of the endometrial receptivity array for repeated implantation failure in Japan: A retrospective, two-centers study. Reprod Med Biol. 2017;16:290–6.

124.Hassold T, Hunt P. Maternal age and chromosomally abnormal pregnancies: what we know and what we wish we knew. Curr Opin Pediatr [Internet] 2009;21(6):703–8.

125.Hassold, T., Hunt, P., 2001. To err (meiotically) is human: the genesis of human aneuploidy. Nature Reviews Genetics 2, 280–291.

126.Hatırnaz Ş, Kanat Pektaş M. Day 3 embryo transfer versus day 5 blastocyst transfers: A prospective randomized controlled trial. J Turkish Soc Obstet Gynecol 2017;14:82–88. Available at: http://cms.galenos.com.tr/Uploads/Article_15859/82-88.pdf.

127.Hashimoto T, Koizumi M, Doshida M, Toya M, Sagara E, Oka N, et al. Efficacy of the endometrial receptivity array for repeated implantation failure in Japan: A retrospective, two-centers study. Reprod Med Biol. 2017;16:290–6.

128.Haug, C.J., Drazen, J.M., 2023. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. New England Journal of Medicine 388, 1201–1208.

129.Hazlina, N.H.N., Norhayati, M.N., Bahari, I.S., Arif, N.A.N.M., 2022. Worldwide prevalence, risk factors and psychological impact of infertility among women: a systematic review and meta-analysis. BMJ Open 12, e057132.

130.He P, Jain R, Ley S, Jacques C, Chambost J, Kotrotsou M, et al. Human labelling of ICSI videos is time-consuming: AI is needed to help embryologists process data they do not have time for. Hum Fertil. 2021;24(1):46–69.

131.Hendriks S, Dancet EA, van Pelt AM, Hamer G, Repping S. Artificial gametes: a systematic review of biological progress towards clinical application. Hum Reprod Update. 2015;21:285–96.

132. Hendrix, N., Veenstra, D.L., Cheng, M., Anderson, N.C., Verguet, S., 2022. Assessing the Economic Value of Clinical Artificial Intelligence: Challenges and Opportunities. Value in Health 25, 331–339.

133.HFEA. Ethnic diversity in fertility treatment. 2021. https:// www.hfea.gov.uk/about-us/publications/ research-and-data/ ethnic-diversity-in-fertility-treatment-2018/

134.HFEA website: https://www.hfea.gov.uk/treatments/explore-all-treatments/intrauterine-insemination-iui/ Last accessed August 2023.

135.HFEA. Fertility treatment 2021: preliminary trends and figures: https://www.hfea.gov.uk/about-us/publications/research-and-data/fertility-treatment-2021-preliminary-trends-and-figures/. Last accessed August 2023.

136.Hickman CFL, Alshubbar H, Chambost J, Jacques C, Pena CA, Drakeley A, et al. Data sharing: using blockchain and decentralized data technologies to unlock the potential of artificial intelligence: What can assisted reproduction learn from other areas of medicine? Fertil Steril. 2020;114:927–33.

137.Hicks SA, Andersen JM, Witczak O, Thambawita V, Halvorsen P, Hammer HL, et al. Machine Learning-Based Analysis of Sperm Videos and Participant Data for Male Fertility Prediction. Sci Rep [Internet] 2019;9(1):16770.

138.Hinojosa, S.M., Hoces, L.N., Guzmán, L., 2022. Time-Lapse Embryo culture: A better understanding of embryo development and clinical application. JBRA Assisted Reproduction. https://doi.org/10.5935/1518-0557.20210107

139.Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q., 2018. Deep learning for image-based cancer detection and diagnosis survey. Pattern Recognition 83, 134 – 149.

140.Huang, L., Bogale, B., Tang, Y., Lu, S., Xie, X.S., Racowsky, C., 2019. Noninvasive preimplantation genetic testing for aneuploidy in spent medium may be more reliable than trophectoderm biopsy. Proceedings of the National Academy of Sciences 116, 14105–14112.

141.Iews, M., Tan, J., Taskin, O., Alfaraj, S., AbdelHafez, F.F., Abdellah, A.H., Bedaiwy, M.A., 2018. Does preimplantation genetic diagnosis improve reproductive outcome in couples with recurrent pregnancy loss owing to structural chromosomal rearrangement? A systematic review. Reproductive BioMedicine Online 36, 677–685.

142.Ishihara O, Arce JC. Japanese Follitropin Delta Phase 3 Trial G. Individualized follitropin delta dosing reduces OHSS risk in Japanese IVF/ICSI patients: a randomized controlled trial. Reproductive biomedicine online. 2021;42:909–18.

143.Iwasaki W, Yamanaka K, Sugiyama D, Teshima Y, Briones-Nagata MP, Maeki M, et al. Simple separation of good quality bovine oocytes using a microfluidic device. Sci Rep. 2018;8:14273.

144.Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:389–91.

145.Jafek A, Feng H, Brady H, Petersen K, Chaharlang M, Aston K, et al. An automated instrument for intrauterine insemination sperm preparation. Sci Rep. 2020;10:21385

146.Jaehwan L, Younsik C, Stephen R, Mun Y, Heon YO, Sungwhan F. Training deep neural networks with 8-bit floating point numbers. 2018; 31: 1–10

147.Jain R, He P, Jacques C, Chambost J, Kotrotsou M, Hickman C. Oolemma response to the needle in ICSI: There is no effect on day 1 outcomes and rates. Hum Fertil. 2021;24(1):46–69.

148.Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 2002;20:422–446.

149.Kanakasabapathy MK, Thirumalaraju P, Kandula H, Doshi F, Sivakumar AD, Kartik D, et al. Adaptive adversarial neural net- works for the analysis of lossy and domain-shifted datasets of medical images. Nat Biomed Eng. 2021;5:571–85.

150.Kaser DJ, Racowsky C. Reply: Clinical outcomes following selection of human preimplantation embryos with time- lapse monitoring: a systematic review. Hum Reprod Update. 2015;21:154.

151.Katz DJ, Teloken P, Shoshany O. Male infertility - the other side of the equation. Aust Fam Physician. 2017;46:641–6.

152.Kaufmann SJ, Eastaugh JL, Snowden S, Smye SW, Sharma V. The application of neural networks in predicting the outcome of in-vitro fertilization. Hum Reprod. 1997;12:1454–7.

153.Khalil M, McGough SA, Pourmirza Z, Pazhoohesh M, Walker S. Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption. A systematic review. Engineering Applications of Artificial Intelligence. 2022;115:105287.

154.Kheradmand et al.2016. Kheradmand, S., Saeedi, P., and Bajic, I. (2016). Human blastocyst segmentation using neural network. In 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pages 1–4. IEEE.

155.Kheradmand et al.2017. Kheradmand, S., Singh, A., Saeedi, P., Au, J., and Havelock, J. (2017). Inner cell mass segmentation in human hmc embryo images using fully convolutional network. In 2017 IEEE International Conference on Image Processing (ICIP), pages 1752–1756. IEEE.

156.Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. npj Digit Med 2019;2:21.

157.Kieslinger DC, De Gheselle S, Lambalk CB, De Sutter P, Kostelijk EH, Twisk JWR, van Rijswijk J, Van den Abbeel E, Vergouw CG. Embryo selection using time-lapse analysis (Early Embryo Viability Assessment) in conjunction with standard morphology: a prospective two-center pilot study. Hum Reprod 2016;31:2450–2457.

158.Kolte AM, Bernardi LA, Christiansen OB, Quenby S, Farquharson RG, Goddijn M, et al. Terminology for pregnancy loss prior to viability: a consensus statement from the ESHRE early pregnancy special interest group. Hum Reprod [Internet] 2015;30(3):495–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25376455

159.Kragh MF, Karstoft H. Embryo selection with artificial intelligence: how to evaluate and compare methods? J Assist Reprod Genet 2021;38(7):1675–89.

160.Kovacs P. Embryo selection: the role of time-lapse monitoring. Reprod Biol Endocrinol [Internet] 2014;12:124.

161.Kumar N, Singh AK. Trends of male factor infertility, an important cause of infertility: a review of literature. J Hum Reprod Sci. 2015;8:191–6.

162.Kuwayama, M., Vajta, G., Ieda, S., Kato, O., 2005a. Comparison of open and closed methods for vitrification of human embryos and the elimination of potential contamination. Reproductive biomedicine online 11, 608–614.

163.Kuwayama M, Vajta G, Ieda S, Kato O. 2005b. Comparison of open and closed methods for vitrification of human embryos and the elimination of potential contamination. Reprod Biomed Online 2005;11(5):608–14.

164.Kuwayama M. Highly efficient vitrification for cryopreservation of human oocytes and embryos: The Cryotop method. Theriogenology 2007;67(1):73–80.

165.Lagalla, C., Barberi, M., Orlando, G., Sciajno, R., Bonu, M. A., Borini, A., May 2015. A quantitative approach to blastocyst quality evaluation: morphometric analysis and related IVF outcomes. Journal of Assisted Reproduction and Genetics 32 (5), 705–712.

166.Laws1980. Laws, K. I. (1980). Textured image segmentation. Technical report, University of Southern California Los Angeles Image Processing INST.

167.LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

168.Lee, C.-I., Wu, C.-H., Pai, Y.-P., Chang, Y.-J., Chen, C.-I., Lee, T.-H., Lee, M.-S., 2019. Performance of preimplantation genetic testing for aneuploidy in IVF cycles for patients with advanced maternal age, repeat implantation failure, and idiopathic recurrent miscarriage. Taiwanese Journal of Obstetrics and Gynecology 58, 239–243.

169.Lee, M., Lofgren, K.T., Thomas, A., Lanes, A., Goldman, R., Ginsburg, E.S., Hornstein, M.D., 2021. The cost-effectiveness of preimplantation genetic testing for aneuploidy in the United States: an analysis of cost and birth outcomes from 158,665 in vitro fertilization cycles. American Journal of Obstetrics and Gynecology 225, 55.e1-55.e17.

170.Le Gac S, Nordhoff V. Microfluidics for mammalian embryo culture and selection: where do we stand now? Mol Hum Reprod. 2017;23:213–26.

171.Lee C-I, Wu C-H, Pai Y-P, Chang Y-J, Chen C-I, Lee T-H, et al. Performance of preimplantation genetic testing for aneuploidy in IVF cycles for patients with advanced maternal age, repeat implantation failure, and idiopathic recurrent miscarriage. Taiwan J Obstet Gynecol [Internet] 2019;58(2):239–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30910146

172.Lee M, Lofgren KT, Thomas A, Lanes A, Goldman R, Ginsburg ES, Hornstein MD. The cost-effectiveness of preimplantation genetic testing for aneuploidy in the United States: an analysis of cost and birth outcomes from 158,665 in vitro fertilization cycles. Am J Obstet Gynecol. 2021;225(1):55.e1–55.e17.

173.Leung C, Lu Z, Esfandiari N, Casper RF, Sun Y. Automated sperm immobilization for intracytoplasmic sperm injection. IEEE Trans Biomed Eng. 2011;58:935–42.

174.Liang B, Gao Y, Xu J, Song Y, Xuan L, Shi T, Wang N, Hou Z, Zhao Y-L, Huang WE, et al. Raman profiling of embryo culture medium to identify aneuploid and euploid embryos. Fertil Steril 2019;111:753-762.e1.

175.Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125.

176.Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988.

177.Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322:1806–16.

178.Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 2008;17:145–151.

179.Loewke K, Cho JH, Brumar CD, Maeder-York P, Barash O, Malmsten JE, Zaninovic N, Sakkas D, Miller KA, Levy M, VerMilyea MD. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. Fertil Steril. 2022 Mar;117(3):528-535.

180.Lu Z, Zhang X, Leung C, Esfandiari N, Casper R, Sun Y. Robotic ICSI (Intracytoplasmic Sperm Injection). IEEE Trans Biomed Eng. 2011;58:2102–8.

181.Ma R, Xie L, Han C, Su K, Qiu T, Wang L, et al. In vitro fertilization on a single-oocyte positioning system integrated with motile sperm selection and early embryo development. Anal Chem. 2011;83:2964–70.

182.Mahadevan MD, Samanta B. Absolute fit: A paradigm shift in quantitative model evaluation in nonlinear mixed effects model framework. Aaps Journals Org 2017;19(3):828–838.

183.Mahadevaiah G, RV P, Bermejo I, Jaffray D, Dekker A, Wee L. Artificial intelligence-based clinical decision support in modern medical physics: selection, acceptance, commissioning, and quality assurance. Med Phys 2020;47: e228–35.

184.Mahajan N. Endometrial receptivity array: Clinical application. J Hum Reprod Sci. 2015;8:121–9.

185.Mahdavi H, Monadjemi A, Vafae A. Sperm detection in video frames of semen sample using morphology and effective ellipse detection method. J Med Signals Sensors 2011;1(3):8.

186.Majumdar, G., Majumdar, A., Verma, I. C., Upadhyaya, K. C., Jan 2017. Relationship between morphology, euploidy and implantation potential of cleavage and blastocyst stage embryos. Journal of Human Reproductive Sciences 10 (1), 49–57.

187.Marti E, de Miguel MA, Garcia F., Perez J. A Review of Sensor Technologies for Perception in Automated Driving. IEEE Intelligent Transportation Systems Magazine. 2019;11(4): 94-108.

188.Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw 2008; 21:427–36.

189.Mendizabal-Ruiz, G., Chavez-Badiola, A., Figueroa, I.A., Nuño, V.M., Farias, A.F.-S., Valencia-Murilloa, R., Drakeley, A., Garcia-Sandoval, J.P., Cohen, J., 2022. Computer software (SiD) assisted real-time single sperm selection associated with fertilization and blastocyst formation. Reproductive BioMedicine Online 45, 703–711.

190.Meseguer M, Herrero J, Tejera A, Hilligsoe KM, Ramsing NB, Remohi J. The use of morphokinetics as a predictor of embryo implantation. Hum Reprod. 2011;26:2658–71.

191.Mio, Y., Mar 2006. Morphological analysis of human embryonic development using time-lapse cinematography. Journal of Mammalian Ova Research 23 (1), 27–35.

192.Mirsky, S.K., Barnea, I., Levi, M., Greenspan, H., Shaked, N.T., 2017. Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning. Cytometry Part A 91, 893–900.

193.Miwa A, Noguchi Y, Hosoya K, Mori Y, Sato T, Kasahara Y, et al. Equivalent clinical outcome after vitrified-thawed blastocyst transfer using semi-automated embryo vitrification system compared with manual vitrification method. Reprod Med Biol. 2020;19:164–70.

194.Mokhtare A, Xie P, Davaji B, Abbaspourrad A, Rosen- waks Z, Palermo G. O-124 Contact-free oocyte denudation in a chip-scale ultrasonic microfluidic device. Hum Reprod. 2021;36(Supplement_1).

195.Munné, S., Lee, A., Rosenwaks, Z., Grifo, J., Cohen, J., 1993. Fertilization and early embryology: Diagnosis of major chromosome aneuploidies in human preimplantation embryos. Human Reproduction 8, 2185–2191.

196.Munné, S., Sultan, K.M., Weier, H.-U., Grifo, J.A., Cohen, J., Rosenwaks, Z., 1995. Assessment of numeric abnormalities of X, Y, 18, and 16 chromosomes in preimplantation human embryos before transfer. American Journal of Obstetrics and Gynecology 172, 1191–1201.

197. Munné S, Kaplan B, Frattarelli JL, Child T, Nakhuda G, Shamma FN, Silverberg K, Kalista T, Handyside AH, Katz-Jaffe M, et al. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. Fertil Steril 2019.

198. Munne S, Kaplan B, Frattarelli JL, Gysler M, Child TJ, Nakhuda G, Shamma FN, Silverberg K, Kalista T, Oliver K, et al. Global multicenter randomized controlled trial comparing single embryo transfer with embryo selected by preimplantation genetic screening using next-generation sequencing versus morphologic assessment. Fertil Steril 2017;108:e19.

199.Mutlag WK, Shaker AK, Zahoor AM, Bahaa TH. Feature Extraction Methods: A Review 2020 J. Phys.: Conf. Ser. 1591 012028.

200.Natesan, S.A., Handyside, A.H., Thornhill, A.R., Ottolini, C.S., Sage, K., Summers, M.C., Konstantinidis, M., Wells, D., Griffin, D.K., 2014. Live birth after PGD with confirmation by a comprehensive approach (karyomapping) for simultaneous detection of monogenic and chromosomal disorders. Reproductive BioMedicine Online 29, 600–605.

201.Nelson SM, Fleming R, Gaudoin M, Choi B, Santo-Domingo K, Yao M. Antimullerian hormone levels and antral follicle count as prognostic indicators in a personalized prediction model of live birth. Fertil Steril. 2015;104:325–32.

202.Nelson SM, Klein BM, Arce JC. Comparison of antimullerian hormone levels and antral follicle count as predictor of ovarian response to controlled ovarian stimulation in good-prognosis patients at individual fertility clinics in two multicenter trials. Fertil Steril. 2015;103:923-30 e1.

203.Neuhausser, W.M., Fouks, Y., Lee, S.W., Macharia, A., Hyun, I., Adashi, E.Y., Penzias, A.S., Hacker, M.R., Sakkas, D., Vaughan, D., 2023. Acceptance of genetic editing and of whole genome sequencing of human embryos by patients with infertility before and after the onset of the COVID-19 pandemic. Reproductive BioMedicine Online 47, 157–163. https://doi.org/10.1016/j.rbmo.2023.03.013

204.Nichol AD, Bailey M, Cooper DJ. Challenging issues in randomised controlled trials. POLAR; EPO Investigators. Injury 2010;41(Suppl 1):S20–3.

205.Niederberger, C., Pellicer, A., Cohen, J., Gartner, D.K., Palermo, G.D., O'Neill, C.L., Chow, S., Rosenwaks, Z., Cobo, A., Swain, J.E., Schoolcraft, W.B., Frydman, R., Bishop, L.A., Aharon, D., Gordon, C., New, E., Decherney, A., Tan, S.L., Paulson, R.J., Goldfarb, J.M., Brännström, M., Donnez, J., Silber, S., Dolmans, M.-M., Simpson, J.L., Handyside, A.H., Munné, S., Eguizabal, C., Montserrat, N., Belmonte, J.C.I., Trounson, A., Simon, C., Tulandi, T., Giudice, L.C., Norman, R.J., Hsueh, A.J., Sun, Y., Laufer, N., Kochman, R., Eldar-Geva, T., Lunenfeld, B., Ezcurra, D., D'Hooghe, T., Fauser, B.C.J.M., Tarlatzis, B.C., Meldrum, D.R., Casper, R.F., Fatemi, H.M., Devroey, P., Galliano, D., Wikland, M., Sigman, M., Schoor, R.A., Goldstein, M., Lipshultz, L.I., Schlegel, P.N., Hussein, A., Oates, R.D., Brannigan, R.E., Ross, H.E., Pennings, G., Klock, S.C., Brown, S., Steirteghem, A.V., Rebar, R.W., LaBarbera, A.R., 2018. Forty years of IVF. Fertility and Sterility 110, 185-324.e5.

206.Nikshad A, Aghlmandi A, Safaralizadeh R, Aghebati-Maleki L, Warkiani ME, Khiavi FM, et al. Advances of microfluidic technology in reproductive biology. Life Sci. 2021;265:118767.

207.Nixon M, Aguado A. Feature extraction and image processing for computer vision. Academic press; 2019 Nov 17.

208.Ogur, C., Kahraman, S., Griffin, D.K., Yapan, C.C., Tufekci, M.A., Cetinkaya, M., Temel, S.G., Yilmaz, A., 2023. PGT for structural chromosomal rearrangements in 300 couples reveals specific risk factors but an interchromosomal effect is unlikely. Reproductive BioMedicine Online 46, 713–727.

209.Okonofua, F.E., Ntoimo, L.F.C., Omonkhua, A., Ayodeji, O., Olafusi, C., Unuabonah, E., Ohenhen, V., 2022. Causes and Risk Factors for Male Infertility: A Scoping Review of Published Studies. International Journal of General Medicine Volume 15, 5985–5997.

210.Ombelet W. Global access to infertility care in developing countries: a case of human rights, equity and social justice. Facts Views Vis Obgyn. 2011;3(4):257-66.

211.Ornes S. AI System Beats Chess Puzzles With 'Artificial Brainstorming'. Nov 2023, Quantum Magazine. Last accessed Nov 18, 2023.

212.Palermo, G., 1992. Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. The Lancet 340, 17–18.

213.Patounakis G, Hill MJ. The preimplantation genetic testing debate continues: first the hype, then the tension, now the hypertension? Fertil Steril 2019;112:233–234.

214.Pennetta F, Lagalla C, Borini A. Embryo morphokinetic characteristics and euploidy. Curr Opin Obstet Gynecol 2018;30:185–96.

215.Phillips-Wren G, Adya M. Decision making under stress: the role of information overload, time pressure, complexity, and uncertainty. J Decis Syst. Published online May 27, 2020.

216.Pirtea P, De Ziegler D, Tao X, Sun L, Zhan Y, Ayoubi JM, et al. Rate of true recurrent implantation failure is low: results of three successive frozen euploid single embryo transfers. Fertil Steril. 2021;115:45–53.

217.Pool TB, Schoolfield J, Han D. Human embryo culture media comparisons. Methods Mol Biol 2012;912:367–86.

218.Popescu F, Jaslow CR, Kutteh WH. Recurrent pregnancy loss evaluation combined with 24-chromosome microarray of miscarriage tissue provides a probable or definite cause of pregnancy loss in over 90% of patients. Hum Reprod 2018;33:579–587.

219.Poynton, C. (2012). Digital video and HD: Algorithms and Interfaces. Elsevier.

220.Practice Committee of the American Society for Reproductive Medicine. Electronic address: asrm@asrm.org. Definitions of infertility and recurrent pregnancy loss: a committee opinion. Fertil Steril. 2020 Mar;113(3):533-535

221.Pribenszky C, Nilselid AM, Montag M. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. Reprod Biomed Online. 2017 Nov;35(5):511-520.

222.Quinn MM, Jalalian L, Ribeiro S, Ona K, Demirci U, Cedars MI, et al. Microfluidic sorting selects sperm for clinical use with reduced DNA damage compared to density gradient centrifugation with swim-up in split semen samples. Hum Reprod. 2018;33:1388–93.

223.Racowsky C, Kovacs P, Martins WP. A critical appraisal of time-lapse imaging for embryo selection: where are we and where do we need to go? J Assist Reprod Genet. 2015;32:1025–30.

224.Rad, R. M., Saeedi, P., Au, J., and Havelock, J. (2017). Coarse-to-fine texture analysis for inner cell mass identification in human blastocyst microscopic images. In 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–5. IEEE.

225.Rad, R. M., Saeedi, P., Au, J., and Havelock, J. (2018a). Human blastocyst's zona pellucida segmentation via boosting ensemble of complementary learning. Informatics in Medicine Unlocked, 13:112–121.

226.Rad, R. M., Saeedi, P., Au, J., and Havelock, J. (2018b). Multi-resolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 3518–3522. IEEE.

227.Rad, R. M., Saeedi, P., Au, J., and Havelock, J. (2019a). Blast-net: Semantic segmentation of human blastocyst components via cascaded atrous pyramid and dense progressive upsampling. In 2019 IEEE International Conference on Image Processing (ICIP), pages 1865–1869. IEEE.

228.Rad, R. M., Saeedi, P., Au, J., and Havelock, J. (2019b). Predicting human embryos' implantation outcome from a single blastocyst image. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 920–924. IEEE.

229.Rad, R. M., Saeedi, P., Au, J., and Havelock, J. (2020). Trophectoderm segmentation in human embryo images via inceptioned u-net. Medical Image Analysis, 62:101612.

230.Rahman NHA, Hasikin K, Razak NAA, Al-Ani AK, Anni DJS, Mohandas P, Medical Device Failure Predictions Through AI-Driven Analysis of Multimodal Maintenance Records. IEEE Access. 2023; 11: 93160-93179.

231.Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. Nature Medicine 28, 31–38.

232.Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. N Engl J Med. 2023 May 25;388(21):1981-1990.

233.Rebouças Filho, P., Rebouças, E., Marinho, L., Sarmento, R., Tavares, J., de Albuquerque, V., 2017. Analysis of human tissue densities: A new approach to extract features from medical images. Pattern Recognition Letters 94, 211–218.

234.Ribeiro, C.C., Camard, J., Silvestri, G., Marques, M., Serrano-Albal, M., Robinson, G., Bradu, A., Chavez-Badiola, A., Podoleanu, A., Griffin, D.K., 2023. P-295 Application of a Time-Lapse Optical Coherence Tomography (OCT) approach in a pilot study to visualise oocytes and embryos in depth. Human Reproduction 38.

235.Riegler MA, Stensen MH, Witczak O, Andersen JM, Hicks SA, Hammer HL, et al. Artificial intelligence in the fertility clinic: status, pitfalls and possibilities. Hum Reprod [Internet] 2021;36(9):2429–42.

236.Riestenberg C, Kroener L, Quinn M, Ching K, Ambartsumyan G. Routine endometrial receptivity array in first embryo transfer cycles does not improve live birth rate. Fertil Steril. 2021;115:1001–6.

237.Rinehart LA. Storage, transport, and disposition of gametes and embryos: legal issues and practical considerations. Fertil Steril. 2021;115:274–81.

238.Rocha, J., Bezerra da Silva, D., dos Santos, J., Whyte, L., Hickman, C., Laver, S., Gouveia Nogueira, M., 2017a. Using artificial intelligence to improve the evaluation of human blastocyst morphology. In: Proceedings of the 9th International Joint Conference on Computational Intelligence - Volume 1: IJCCI, INSTICC, SciTePress, pp. 354–359.

239.Rocha, J., Passalia, F., Matos, F., Takahashi, M., Ciniciato, D. d. S., Maserati, M., Alves, M., de Almeida, T., Cardoso, B., Basso, A., Nogueira, M., 2017b. A method based on artificial intelligence to fully automatize the evaluation of bovine blastocyst images. Scientific Reports 7 (1), 7659.

240.Rocha, J., Passalia, F., Matos, F., Takahashi, M., Maserati, M., Alves, M., de Almeida, T., Cardoso, B., Basso, A., Nogueira, M., 2017c. Automatized image processing of bovine blastocysts produced in vitro for quantitative variable determination. Scientific Data 4, 170192.

241.Roque M, Simon C. Time to pregnancy: as important for patients as underestimated by doctors. Fertil Steril [Internet] 2020;113(3):522–3.

242.Rooney KL, Domar AD. The relationship between stress and infertility. Dialogues Clin Neurosci [Internet] 2018;20(1):41–7.

243.Rose, N.C., Kaimal, A.J., Dugoff, L., Norton, M.E., American College of Obstetricians and Gynecologists, 2020. Screening for fetal chromosomal abnormalities: ACOG practice bulletin, number 226. Obstetrics & Gynecology 136, e48–e69.

244.Rosenwaks Z, Handyside AH, Fiorentino F, Gleicher N, Paulson RJ, Schattman GL, et al. The pros and cons of preimplantation genetic testing for aneuploidy: clinical and laboratory perspectives. Fertil Steril. 2018;110:353–61.

245.Rosenwaks, Z., 2020. Artificial intelligence in reproductive medicine: a fleeting concept or the wave of the future? Fertility and Sterility 114, 905–907.

246.Roy TK, Brandi S, Tappe NM, Bradley CK, Vom E, Henderson C, et al. Embryo vitrification using a novel semi-automated closed system yields in vitro outcomes equivalent to the manual Cryotop method. Human Reprod (Oxford, England). 2014;29:2431–8.

247.Rubio C, Bellver J, Rodrigo L, Castillón G, Guillén A, Vidal C, et al. In vitro fertilization with preimplantation genetic diagnosis for aneuploidies in advanced maternal age: a randomized, controlled study. Fertil Steril [Internet] 2017;107(5):1122–9.

248.Rubio C, Navarro-Sanchez L, Garcia-Pascual CM, Ocali O, Cimadomo D, Venier W, et al. Multicenter prospective study of concordance between embryonic cell-free DNA and trophectoderm biopsies from 1301 human blastocysts. Am J Obstet Gynecol. 2020;223(751):e1–13.

249.Ruiz-Alonso M, Blesa D, Diaz-Gimeno P, Gomez E, Fernandez- Sanchez M, Carranza F, et al. The endometrial receptivity array for diagnosis and personalized embryo transfer as a treatment for patients with repeated implantation failure. Fertil Steril. 2013;100:818–24.

250.Ryman-Tubb NF, Krause P, Garn W. How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. Engineering Applications of Artificial Intelligence. 2018; 76: 130-157.

251.Saadat M, Hajiyavand AM, Singh Bedi AP. Oocyte Positional Recognition for Automatic Manipulation in ICSI. Micromachines (Basel). 2018;9(9):429.

252.Saito, T., Rehmsmeier, M., 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10, e0118432.

253.Sahni NR, Carrus B. Artificial Intelligence in U.S. Health Care Delivery. N Engl J Med. 2023; 27;389(4):348-358.

254.Salih M, Austin C, Warty RR, Tiktin C, Rolnik DL, Momeni M, Rezatofighi H, Reddy S, Smith V, Vollenhoven B, Horta F. Embryo selection through artificial intelligence versus embryologists: a systematic review. Hum Reprod Open. 2023; 15;2023(3)

255.Sánchez D, González A, Flores-Saiffe R, Valencia-Murillo G, Mendizabal-Ruiz A, Chavez-Badiola. P–245 Machine learning predicting oocyte's fertilization and blastocyst potential based on morphological features. Hum Reprod. 2021;36(Supplement_1).

256.Sanders KD, Silvestri G, Gordon T, Griffin DK. Analysis of IVF live birth outcomes with and without preimplantation genetic testing for aneuploidy (PGT-A): UK Human Fertilisation and Embryology Authority data collection 2016-2018. J Assist Reprod Genet. 2021; 38(12):3277-3285.

257.Sanson-Fisher RW, Bonevski B, Green LW, D'Este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. Am J Prev Med 2007;33:155–61.

258.SART. National Summary Report. 2018. https://www.sartc orsonline.com/rptCSR_PublicMultYear.aspx?reportingYear= 2019.

259.Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evol Intell. 2022;15(1):1-22.

260.Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.

261.Sciorio, R., Thong, J. K., Pickering, S. J., Mar 2018. Comparison of the development of human embryos cultured in either an embryoscope or benchtop incubator. Journal of Assisted Reproduction and Genetics 35 (3), 515–522.

262.Sciorio, R., Meseguer, M., 2021. Focus on time-lapse analysis: blastocyst collapse and morphometric assessment as new features of embryo viability. Reproductive BioMedicine Online 43, 821–832.

263.Sciorio, R., Rinaudo, P., 2023. Culture conditions in the IVF laboratory: state of the ART and possible new directions. Journal of Assisted Reproduction and Genetics 40, 2591–2607.

264.Scott L. Pronuclear scoring as a predictor of the embryo development. Reprod Biomed Online. 2003;6:201–14.

265.Sharp TA, Garbarini Jr. WN, Johnson CA, Watson A, Go KJ. Proof of concept for an automated tank storing frozen embryos and gametes in an ART laboratory. ASRM: Fertility Sterility. 2019;112(3):Supplement E432.

266.Sharp TA, Garbarini Jr. WN, Johnson CA, Watson A, Greenberg R, Go KJ. Initial validation of an automated cryostorage and inventory management system. ASRM: Fertil Steril. 2019;112(3):Supplement E116.

267.Simon C, Gomez C, Ruiz M, Mol BW, Valbuena D. Response to: Comments on the methodology of an endometrial receptivity array trial. Reprod Biomed Online. 2021;42:284.

268.Steptoe PC. Edwards. RG (1978) Birth after the reimplantation of a human embryo. Lancet.;2(8085):366.

269.Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010 Jan;21(1):128-38.

270.Storr, A., Venetis, C. A., Cooke, S., Susetio, D., Kilani, S., Ledger, W., Jul 2015. Morphokinetic parameters using time-lapse technology and day 5 embryo quality: a prospective cohort study. Journal of Assisted Reproduction and Genetics 32 (7), 1151–1160.

271.Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 12, e1001779.

272.Swain J, VerMilyea MT, Meseguer M, Ezcurra D, Fertility AIFG. AI in the treatment of fertility: key considerations. J Assist Reprod Genet. 2020;37:2817–24.

273.Tang, Q., Liu, Y. & Liu, H. Medical image classification via multiscale representation learning. Artif. Intell. Medicine 79, 71–78.

274.Theobald R, SenGupta S, Harper J. The status of preimplantation genetic testing in the UK and USA. Hum Reprod. 2020;35:986–98.

275.Thomas D, Flanagan J, Monteiro M, Maillot M, Simon Z, Taha M, et al. Clinical evaluation of artificial intelligence robotic tool for male factor infertility. Hum Fertil. 2021;24(1):46–69.

276.Tıras B, Cenksoy PO. Practice of embryo transfer: recommendations during and after. Semin Reprod Med [Internet] 2014;32(4):291–6.

277.Tong, J., Niu, Y., Wan, A., Zhang, T., 2022. Effect of parental origin and predictors for obtaining a euploid embryo in balanced translocation carriers. Reproductive BioMedicine Online 44, 72–79.

278.Tosti, E., Ménézo, Y., 2016. Gamete activation: basic knowledge and clinical applications. Human Reproduction Update 22, 420–439. https://doi.org/10.1093/humupd/dmw014

279.Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, La VP, Ho MT, Nghiem KP, Nguyen HLT, Latkin CA, Tam WWS, Cheung NM, Nguyen HT, Ho CSH, Ho RCM. Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. J Clin Med. 2019 Mar 14;8(3):360.

280.Tran D, Cooke S, Illingworth PJ, Gartner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Hum Reprod 2019;34:1011–1018.

281.Treff, N.R., Eccles, J., Marin, D., Messick, E., Lello, L., Gerber, J., Xu, J., Tellier, L.C.A.M., 2020. Preimplantation Genetic Testing for Polygenic Disease Relative Risk Reduction: Evaluation of Genomic Index Performance in 11,883 Adult Sibling Pairs. Genes 11, 648.

282.Treff, N.R., Zimmerman, R., Bechor, E., Hsu, J., Rana, B., Jensen, J., Li, J., Samoilenko, A., Mowrey, W., Alstine, J.V., Leondires, M., Miller, K., Paganetti, E., Lello, L., Avery, S., Hsu, S., Tellier, L.C.A.M., 2019. Validation of concurrent preimplantation genetic testing for polygenic and monogenic disorders, structural rearrangements, and whole and segmental chromosome aneuploidy with a single universal platform. European Journal of Medical Genetics 62, 103647.

283.Ulahannan TJ. Decision Making in Health and Medicine: Integrating Evidence and Values. J R Soc Med. 2002 Feb;95(2):108–9. PMCID: PMC1279329.

284.Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14:e0224365.

285.Vander Borght M, Wyns C. Fertility and infertility: definition and epidemiology. Clin Biochem. 2018;62:2–10.

286.Van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis. 2022;79:102470.

287.VerMilyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, et al. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Hum Reprod. 2020;35:770–84.

288.Verpoest W, Staessen C, Bossuyt PM, Goossens V, Altarescu G, Bonduelle M, Devesa M, Eldar-Geva T, Gianaroli L, Griesinger G, et al. Preimplantation genetic testing for aneuploidy by microarray analysis of polar bodies in advanced maternal age: a randomized clinical trial. Hum Reprod 2018;33:1767–1776.

289.Viotti M, Greco E, Grifo JA, Madjunkov M, Librach C, Cetinkaya M, Kahraman S, Yakovlev P, Kornilov N, Corti L, Biricik A, Cheng EH, Su CY, Lee MS, Bonifacio MD, Cooper AR, Griffin DK, Tran DY, Kaur P, Barnes FL, Zouves CG, Victor AR, Besser AG, Madjunkova S, Spinella F. Chromosomal, gestational, and neonatal outcomes of embryos classified as a mosaic by preimplantation genetic testing for aneuploidy. Fertil Steril. 2023 Nov;120(5):957-966.

290.Wang, R., Pan, W., Jin, L., Li, Y., Geng, Y., Gao, C., Chen, G., Wang, H., Ma, D., Liao, S., 2019. Artificial intelligence in reproductive medicine. Reproduction 158, R139–R154.

291.Wang S, Summers RM. Machine learning and radiology. Med Image Anal 2012;16:933–51.

292.Watanabe Y, Miura K, Kumagai Y, Matsumoto M, Noda T, Yoshimura Y. Supervised learning for morphological analysis of human embryos. Comput Biol Med. 2013;43(11):1582-7.

293.Wei L, Zhang J, Shi N, Luo C, Bo L, Lu X, Gao S, Mao C. Association of maternal risk factors with fetal aneuploidy and the accuracy of prenatal aneuploidy screening: a correlation analysis based on 12,186 karyotype reports. BMC Pregnancy Childbirth. 2023 Mar 2;23(1):136.

294.Weissman A, Shoham G, Shoham Z, Fishel S, Leong M, Yaron Y. Preimplantation genetic screening: results of a worldwide web-based survey. Reprod Biomed Online 2017;35:693–700.

295.Weng L, Lee GY, Liu J, Kapur R, Toth TL, Toner M. On-chip oocyte denudation from cumulus-oocyte complexes for assisted reproductive therapy. Lab Chip. 2018;18:3892–902.

296.Werbos P. Beyond regression: new tools for prediction and analysis in the behavioral sciences, PhD diss. Harvard University; 1974.

297.Widder, D.G., Nafus, D., Dabbish, L., Herbsleb, J., 2022. Limits and Possibilities for “Ethical AI” in Open Source: A Study of Deepfakes, in: 2022 ACM Conference on Fairness, Accountability, and Transparency. ACM, pp. 2035–2046.

298.Wong FS, Wang PZ, Goh TH, Quek BK. Fuzzy Neural Systems for Stock Selection. Financial Analysts Journal 48, no. 1 (1992): 47–74.

299.Xiao S, Riordon J, Simchi M, Lagunov A, Hannam T, Jarvi K, et al. FertDish: microfluidic sperm selection-in-a-dish for intra- cytoplasmic sperm injection. Lab Chip. 2021;21:775–83.

300.Yang P, Wang Y, Wu Z, Pan N, Yan L, Ma C. Risk of miscarriage in women with endometriosis undergoing IVF fresh cycles: a retrospective cohort study. Reprod Biol Endocrinol [Internet] 2019;17(1):21.

301.Yoshida, I. H., Santos, M., Berton, C. Z., Chiarella, C. L., Tanada, M. S., Cordts, E. B., de Carvalho, W. P., Barbosa, C. P., Apr 2018. Can trophectoderm morphology act as a predictor for euploidy? JBRA Assist Reprod 22 (2), 113–115.

302.Yovich JL, Alsbjerg B, Conceicao JL, Hinchliffe PM, Keane KN. PIVET rFSH dosing algorithms for individualized controlled ovarian stimulation enables optimized pregnancy productivity rates and avoidance of ovarian hyperstimulation syndrome. Drug Des Devel Ther. 2016;10:2561–73.

303.Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng [Internet] 2018;2(10):719–31.

304.Zegers-Hochschild, F., Adamson, G.D., Mouzon, J. de, Ishihara, O., Mansour, R., Nygren, K., Sullivan, E., Poel, S. van der, Technology, I.C. for M.A.R., Organization, W.H., 2009. The International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) Revised Glossary on ART Terminology, 2009. Human reproduction (Oxford, England) 24, 2683–7.

305.Zhang, J., Liu, H., Luo, S., Lu, Z., Chávez-Badiola, A., Liu, Z., Yang, M., Merhi, Z., Silber, S.J., Munné, S., Konstantinidis, M., Wells, D., Tang, J.J., Huang, T., 2017. Live birth derived from

oocyte spindle transfer to prevent mitochondrial disease. Reproductive biomedicine online 34, 361–368.

306.Zhang, J.J., Merhi, Z., Yang, M., Bodri, D., Chavez-Badiola, A., Repping, S., Wely, M. van, 2016. Minimal stimulation IVF vs conventional IVF: a randomized controlled trial. American Journal of Obstetrics and Gynecology 214, 96.e1-96.e8.

307.Zhang WY, von Versen-Höynck F, Kapphahn KI, Fleischmann RR, Zhao Q, Baker VL. Maternal and neonatal outcomes associated with trophectoderm biopsy. Fertil Steril 2019;112:283-290.e2.

308.Zhang Z, Dai C, Shan G, Chen X, Liu H, Abdalla K, et al. Quantitative selection of single human sperm with high DNA integrity for intracytoplasmic sperm injection. Fertil Steril [Internet] 2021