The use of machine learning in assessing mammalian pre-implantation embryo quality

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in in the Faculty of Biology, Medicine, and Health

2023

Camilla A. Mapstone

Division of Developmental Biology and Medicine

[Blank Page]

List of contents

Word count: 40804 words

List of abbreviations

AUC: Area under curve

CNN: convolutional neural network

DNN: deep neural network

DOHaD: developmental origins of health and disease

EGA: embryonic genome activation

EPI: epiblast

ExE: extraembryonic ectoderm

FD: first division

ICM: inner cell mass

ICSI: Intracytoplasmic sperm injection

IVF: In vitro fertilisation

LB: live birth

ML: machine learning

NEBD: nuclear envelope breakdown

NN: neural network

NP: no prenancy

NPB: nucleolus precursor bodies

PCA: principle component analysis

PN: pronuclei

PrE: Primitive endoderm

ROC: receiver operator characteristic

SET: single embryo transfer

t-SNE: t-distributed stochastic neighbour embedding

TE: trophectoderm

TLM: time-lapse monitoring

VE: visceral endoderm

ZP: zona pellucida

List of figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4

List of tables

Chapter 1:

Chapter 2:

Chapter 3:

Abstract

In Vitro Fertilisation (IVF) is quickly becoming an extremely important medical intervention as the prevalence of infertility increases. Therefore, it is vital to ensure that IVF procedures are as safe and successful as possible. There are still many challenges to be addressed, including the embryo assessment process used to select an embryo for transfer. This is a difficult task, as knowledge of pre-implantation development is incomplete and there is still a high degree of subjectivity involved in assessing embryos for viability.

In this thesis, we work towards more accurate, objective and versatile embryo selection by employing machine learning (ML) techniques and investigating a range of potential morphological quality markers, including currently neglected sub-cellular features. We first present CNN models trained to predict live birth from a variety of developmental stages. These include the first DL models predicting live birth using solely pre-blastocyst stages, which could allow for earlier embryo transfer, mitigating the harmful effects of prolonged culture. We also showed that information from earlier stages can assist selection at blastocyst stage, allowing for the previously unachievable ranking of high-quality blastocysts. In developing these models, we explored the time period of pre-implantation development to identify the best developmental moments for predicting live birth, therefore providing crucial information for embryo assessment procedures.

In order to achieve the best possible assessment of embryo quality at pre-blastocyst stage, we next identified morphological features correlated with transfer outcome and combined these with the CNN model outputs to get an improved overall prediction of live birth. We experimented with different supervised learning techniques and found that linear regression gave the best performance. We also investigated the structure of our dataset via dimensionality reduction techniques and unsupervised clustering, gaining a deeper insight into the challenge of embryo selection.

Finally, we investigated changes in nuclear size and appearance during preimplantation development, a sub-cellular feature not currently used to assess embryo viability past the first two embryonic cycles. For this we used the mouse embryo, which mirrors key developmental stages in human embryos. We discovered trends in size and shape both over development, and across different lineages at the same developmental stage. This research could pave the way for better understanding of standard nuclei appearance during pre-implantation development, allowing for the existing embryo selection criteria to be extended.

In this thesis we have demonstrated the potential of ML techniques to increase knowledge of the pre-implantation development period, and ultimately lead to improved embryo selection procedures in IVF.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements

First of all, I would like to thank my supervisors for their enduring support throughout my PhD. My primary supervisor Dr Berenika Plusa, whose infectious enthusiasm has continuously inspired me. Thank you for your patience in introducing me to the field of developmental biology and for teaching me so many new skills. Prof. Julia Handl, whose knowledge and guidance has assisted me through all stages of this PhD and helped me develop as a researcher. And Prof. Daniel Brison, whose expertise and advice has been invaluable throughout the PhD. I feel very fortunate to have had such a supportive supervisory team.

I was very grateful to be financially supported by the Wellcome Trust, who generously funded my research as part of the 'Quantitative and Biophysical Biology' PhD program.

I would like to thank my Mum, Dad, and sisters for their encouragement in pursuing this PhD and their continuous support over the past four and a half years. I am very thankful to Alex for being my emotional support, your understanding helped me through the tough times. Finally, I would like to thank all my close friends for their continuous encouragement and inspiration over the course of this PhD.

Preface

This thesis has been written in journal format, each of the results chapters are presented as self-contained research papers. The manuscript in chapter 2 has been recently been submitted to npj Digital Medicine and subsequently sent to reviewers, and the manuscripts in chapters 3 and 4 will also be submitted for publication in the coming year. This choice of thesis format has enabled me to prepare manuscripts suitable for publication as well as inclusion in the thesis.

My undergraduate degree was an MPhys in Physics at Bristol University, where I gained a broad range of skills that have been applied throughout this PhD program, including programming, data analysis, and communicating results both verbally and in writing. This thesis describes my primary research experience.

Chapter 1:

Introduction

In Vitro Fertilisation (IVF) has now been used for over 40 years, beginning with the birth of Louise Brown in 1978[1], and has become an established practice used to treat infertility with 69,000 treatment cycles in the UK in 2019[2]. The selection of an embryo suitable for transfer is currently carried out by visual assessment of time-lapse videos. This is subjective and can lead to different results between embryologists[3]. Techniques from pattern recognition and machine learning may provide suitable tools to help support this task, leading to more reliable and consistent predictions and the ultimate selection of embryos with a higher chance of a successful pregnancy and healthy birth[4]. For this PhD project I have used both supervised and unsupervised learning methods to examine Bright Field and fluorescence images of pre-implantation mammalian embryos to attempt to make predictions about viability and to gain a greater understanding of pre-implantation development.

Here, I will review the areas of biology, IVF, and machine learning relevant to the research I have carried out whilst studying for my PhD. This will be split into several sections; section 1 will cover current knowledge of pre-implantation embryo development in mammals with a focus in the human and the mouse, section 2 will look at IVF techniques and procedures and go into detail of current practices and limitations, section 3 will look at machine learning techniques, focusing on those that may be useful for embryo selection, and finally section 4 will put all this knowledge into context with the aims of my research.

1. Pre-implantation development

Mammalian pre-implantation development is the transformation of a fertilised egg into an organised structure referred to as the blastocyst. During this development three distinct and separated cell lineages emerge; the trophectoderm (TE), the epiblast (EPI) and the primitive endoderm (PrE)[5]. Each lineage is essential for the continued development of the embryo and a successful pregnancy, the TE gives rise to the foetal part of the placenta, the EPI is the precursor of somatic cells and germ cells of the body and the PrE gives rise to the yolk sac endoderm[5]. A greater understanding of the ability of the embryo to organise itself into these 3 lineages could have important implications for assisted reproductive technology[6].

1.1 Developmental stages in human embryos

The pre-implantation development period begins with fertilization in the oviduct and the resulting single cell embryo, referred to as the zygote, passes through several stages before implanting in the endometrium about a week later[6], as illustrated in figure 1. In the first day after fertilisation the embryo remains at zygote stage and should contain two pronuclei (PN)[6] - transformed sperm and egg chromatin forming two, separate, nucleus like structures (See figure 1A). Both male (delivered by sperm) and female (from the oocyte) chromatin have to undergo substantial chromatin architectural remodelling to form the PN. The two PN move towards each other, meet near the centre of the cell, and then undergo nuclear envelope breakdown shortly before the first division[7]. Within each PN several smaller round bodies can be found, these are referred to as nucleolus precursor bodies[8] (NPB) and are visible in human zygotes (as can be seen in figure 1A).

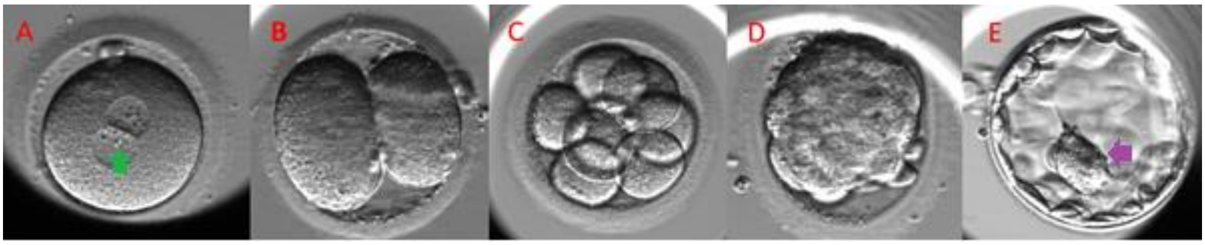


Figure 1: Human Pre-implantation development. A) Zygote stage – the egg has recently been fertilised and the two pronuclei can be observed (NPB indicated by green arrow), B) 2 cell stage, C) Cleavage stage- the cells get progressively smaller as they divide, D) Morula – the embryos compacts and the cells become indistinguishable from each other, E) Blastocyst – A hollow structure forms with the TE cells around the outside and the ICM cells (indicated by purple arrow) located inside the TE, on one side of the cavity. The figure was produced using time-lapse images from St Marys IVF clinic, Manchester

After around 24 hours the zygote begins to undergo cleavages, a special type of cell division where there is no growth stage and the cells get progressively smaller with each cycle[6]. The cells in the embryo are easily distinguishable from each other (figure 1C) until the 8-16 cell stage, occurring at about day 4[6], when the embryo forms a compacted morula[9] (figure 1D). The compacted morula cells are tightly packed together and connected by intercellular junctions that can be detected from the 8 cell stage onwards, but become more elaborate by the 26 cell stage[10]. This compaction stage and formation of the morula is preceded by embryonic genome activation (EGA), the process in which transcription of the embryonic genome begins. Although the first major wave of EGA is usually observed at the 4-8 cell stage in humans, it has been shown that it is independent of cell number, normally occurring at day 3[11].

On day 5 a cavity appears and starts to grow as the blastocyst (a hollow fluid filled structure) is formed. By now the inner cell mass (ICM) and TE lineages have been established and separated; the TE cells form the outer layer of the blastocyst while the ICM cells (precursors or the PrE and EPI) cluster together as a ball inside the cavity (figure 1E). On day 6[6] the embryo hatches from the zona pellucida (ZP), a glycoprotein coat surrounding the embryo, and then is ready to implant in the endometrium on day 7, just after the ICM further differentiates into the EPI and PrE[6] (the latter is often referred to as the hypoblast in human embryos).

1.2 Mouse development

Research into human embryo development is restricted both due to limited supply of good quality embryos available for observation and the restrictions on experimentation for ethical reasons. Therefore, a lot of knowledge about mammalian pre-implantation embryos has come from studying other mammals. So far, this has predominately been the mouse due to relatively low neurosensitivity (lower cognitive abilities and reactivity to stimulus), low costs, and a well developed in vitro culture system[12].

One of the most important questions that research into pre-implantation development attempts to answer is how the first three separate lineages form; there have been many studies looking into the properties of the cells in early mice embryos in an attempt to identify causes or markers of differentiation into a certain lineage. It is known that the early cells (2-8 cell stage) are totipotent, meaning that they can give rise to any of the embryonic and extra-embryonic lineages, but this totopotiency is gradually lost as each lineage specifies[13]. This loss of totopotiency and initiation of cell specification involves two cell fate decisions; TE versus ICM and then EPI versus PrE within the ICM.

TE versus ICM cell fate decision

The first cell fate decision involves the allocation of blastomeres to either the ICM or TE and occurs during the 8-32 cell stage, prior to cavitation[13]. However, it is thought that the full developmental potential of the inside cells is not lost until the early blastocyst stage as it is possible for isolated ICM cells to form structures resembling the blastocyst[14]. At the 8 cell stage the cells of the mouse embryo develop an apical – basal polarity. The subsequent divisions then can produce two different types of cells; symmetrical division leads to two polar daughter cells as both inherit the apical domain, while asymmetrical division results in one outer polar cell and one inner apolar cell. This creates an inner group of apolar cells surrounded by an outer layer of polar cells[13].

The TE and ICM express different levels of lineage-specific transcription factors; this difference is now widely accepted to be a result of the cell polarity status[13]. Lineage specific transcription factor dynamics include the expression of *Cdx2* and Gata3 in the TE precursor cells[15,16], and elevated expression of *Nanog* and *Oct4* and Sox2 in the ICM cells.[17-20]

C*dx2* expression is regulated by the transcription factor TEAD4 via the Hippo pathway (illustrated in figure 2), which if active indirectly prevents TEAD4 from causing *Cdx2* to be expressed[21-25]. The hippo pathway is inactive in the outer cells, leading to greater expression of *Cdx2* and therefore the formation of TE precursor cells, and active in the inner cells leading to a down-regulation of *Cdx2* expression and formation of ICM precursor cells[23-29].

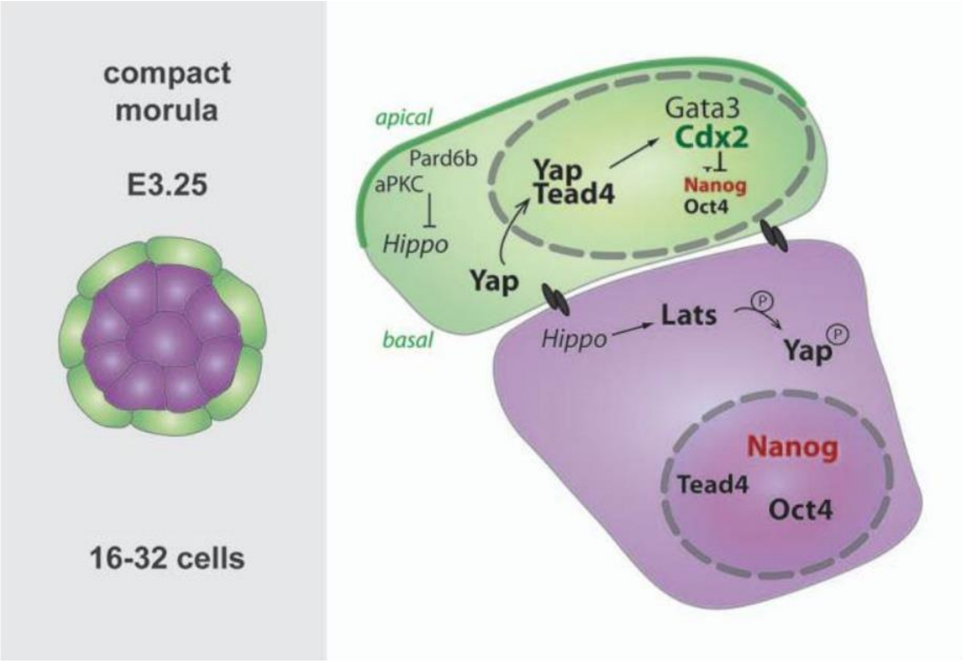


Figure 2: An illustration of Hippo signalling in the early mouse embryo. The inner cells are unpolarised so Hippo signalling is active and causes phosphorylation of YAP. This means that YAP cannot travel into the nucleus where it co-activates TEAD4. In the outer cells Hippo signalling is deactivated due to the presence of polarised apical domain therefore YAP is able to travel into the nucleus and activate TEAD4, leading to the expression of *Cdx2*[13] [11]. Image from Schrode et al. (2013)[13].

EPI versus PrE cell fate decision

The 2nd cell fate decision occurs in the ICM of a developing blastocyst after the 32 cell stage. It results in the PrE emerging as a layer of epithelial cells on the surface of the ICM next to the blastocyst cavity and the rest of the ICM cells forming the EPI[30,31]. The EPI cells are marked by transcription factors associated with pluriopotency; NANOG, SOX2 and OCT4, while the PrE is marked by the transcription factors such as GATA4 and GATA6[17-20,32,33]. There are thought to be three stages in the 2nd cell fate decision[34]:

1. Co-expression of lineage specific transcription factors at and before the 32 cell stage
2. Mutually exclusively expression of transcription factors appearing at around the 64 cell stage with a salt and pepper distribution of the EPI and PrE precursor cells
3. Sorting and segregation of the 2 cell separate lineages.

Live imaging has revealed that there are multiples mechanisms leading to the segregation of the PrE and the EPI once the salt and pepper distribution is established[34]. These include selective apoptosis, downregulation of PrE genes in cells that remain inside ICM and migration of PrE cells from the centre of the ICM to the surface[23] .

1.3 Comparison of Mouse and Human

The pre-implantation mouse embryo passes through the same main stages as the human embryo and is morphologically similar, however there are some key differences. A major difference is in timing and scale; the human embryo reaches compaction and initiates blastocyst formation at a later time than the mouse embryo, and also usually undergoes extra rounds of cell division before implantation[6] (see figure 3). Differences have also been observed in the timing and localisation of the lineage-specific transcription factors throughout the pre-implantation period[12].

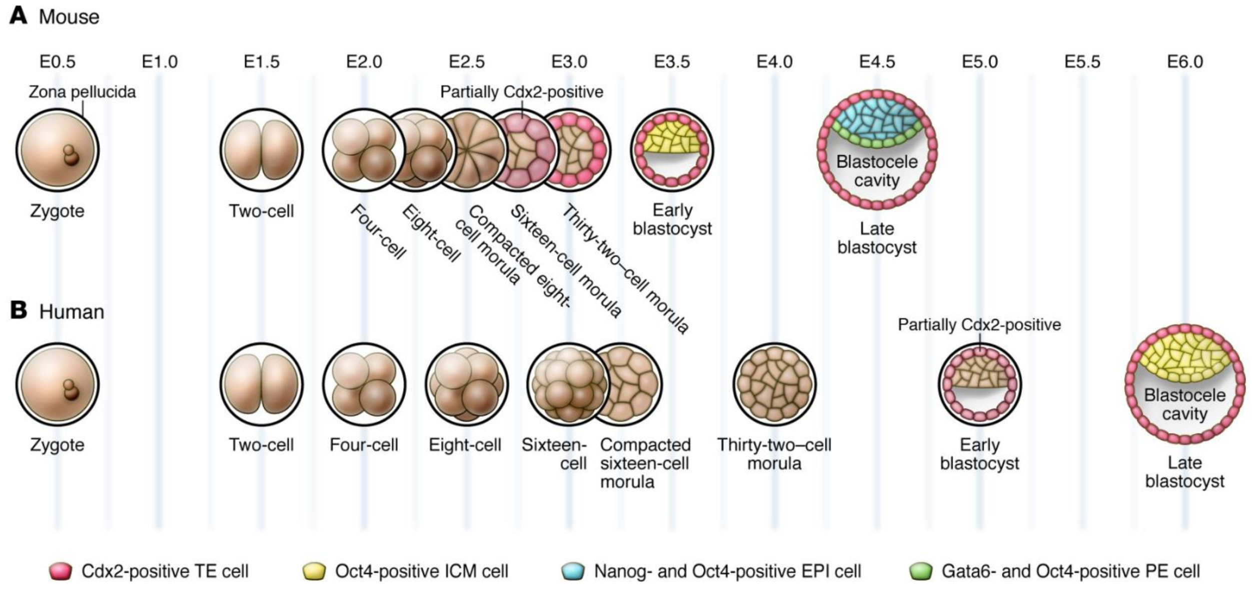


Figure 3: Comparison of pre-implantation development for mouse and human embryos. From Cockburn et al. (2010)[35]

One such difference is in the timing of the localization of CDX2; in human embryos CDX2 does not appear until the blastocyst stage[36], suggesting that it is not involved in the first cell fate decision for human embryos. Conversely, WNT3 and B-CATENIN have been suggested to play a role in the formation of the TE in human embryos[37], despite not having been found to have the same impact at this stage in mouse embryos.

The key transcription factors involved in the second cell fate decision in the mouse embryo are also found in the human embryo but the timings of expression may differ. In human embryos localized SOX2 has been observed as early as the compacted morula, whilst localized NANOG has not been observed until later on in the blastocyst development[38]. GATA6 has also been found to remain present in the TE at a later stage in human embryos than in mice[12]. Despite these differences, by the end of the pre-implantation period the localization of all key lineage- specific transcription factors is the same in human and mice embryos.

2. Assisted reproduction

2.1 Overview

Due to growing problems with infertility, IVF has become an important medical intervention. The number of fertility treatments is increasing (see figure 4), which may be partly due to the general population delaying parenthood as fertility declines with age[39]. As shown in figure 4, there has also been an increase in the number of patients undergoing IVF treatments for non-fertility reasons, such as those in same-sex couples or without a partner, although these cases still make up less than 10% of IVF patients[39].

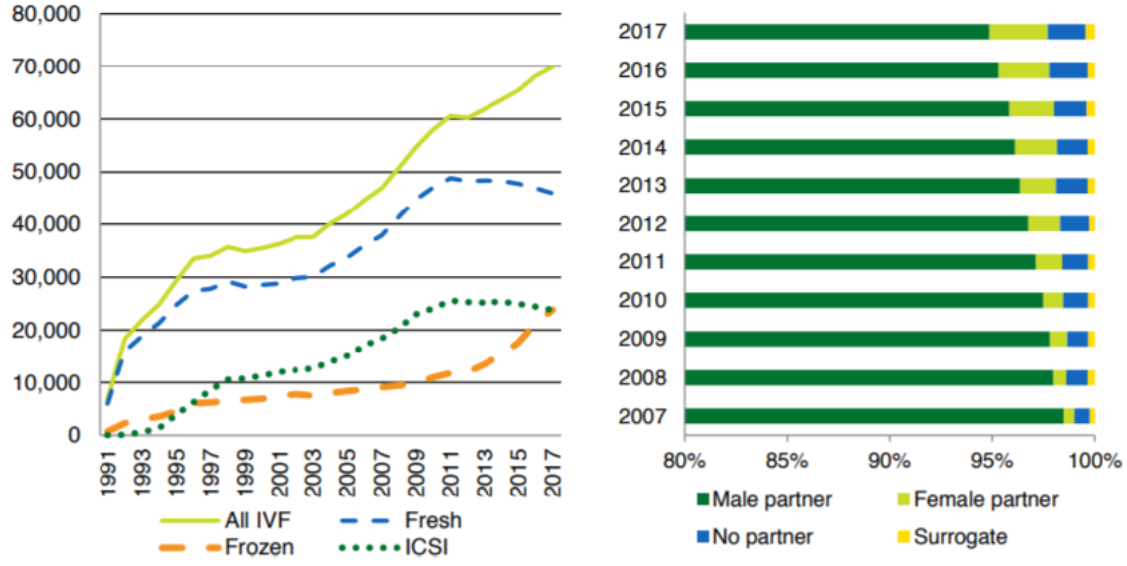


Figure 4: IVF statistics. On the left is the number of various types of IVF procedures in the UK from 1991 to 2017 (ICSI, fresh eggs, frozen eggs and total number), and on the right is the partner status of IVF patients from 2007-2017. Both taken from HFEA[39].

IVF involves collecting the oocytes from the woman’s ovaries to be artificially fertilised. Any resulting embryos are cultured for up to 5 days, then one or more embryos are selected to be transferred back to the woman’s uterus. The eggs can be fertilized either by mixing the oocytes and sperm in a culture dish (traditional IVF) or by using a technique called intracytoplasmic sperm injection (ICSI) where the sperm head is injected directly into the oocyte[40].

Embryo culture is usually extended to blastocyst stage to aid the embryo selection process, as blastocyst culture is correlated with higher live birth rate per embryo[41]. However, the cumulative live birth rate once all embryos in a cycle are transferred is not increased by blastocyst culture[41]. There are also various health risks associated with blastocyst culture; mono-zygotic twinning, reduced birth weight, pre-term birth, and shortened telomeres[41-44]. The recent accumulation of evidence of the risks of prolonged culture may require a re-evaluation of the risk–benefit profile of blastocyst culture[45]

Although there has been significant progress in IVF techniques over the last few decades, success rates still remain low. Chances of a successful pregnancy vary both with age[46] and the number of embryos transferred[47]. In the 1980s, published records of pregnancy rates showed that a pregnancy was significantly more likely when two of more embryos were transferred[47], and the transfer of multiple embryos became common in IVF procedures to maximise success rates[48]. However, multiple gestations have been shown to have a number of risks such as preterm births, congenital abnormalities[48], and increased short term and long term health risks to the mother[49]. This also has an effect on the delivery related expenses for hospitals; the delivery of twins has been shown to cost over twice the amount of the delivery of a singleton per baby[50]. Therefore, in recent years there has been a move towards single embryo transfer (SET; see figure 5), helped in part by educating patients on the risks of multiple gestation[48].

This move towards SET means that it is even more important to select the embryo with the best chance of developing into a healthy baby. This has proved to be challenging, although there is a set of morphological criteria that are used in most IVF clinics there is currently no marker fully predictive of a healthy development[51]. Progress in this area is currently limited by our lack of understanding of pre-implantation development and what constitutes a healthy embryo[12].

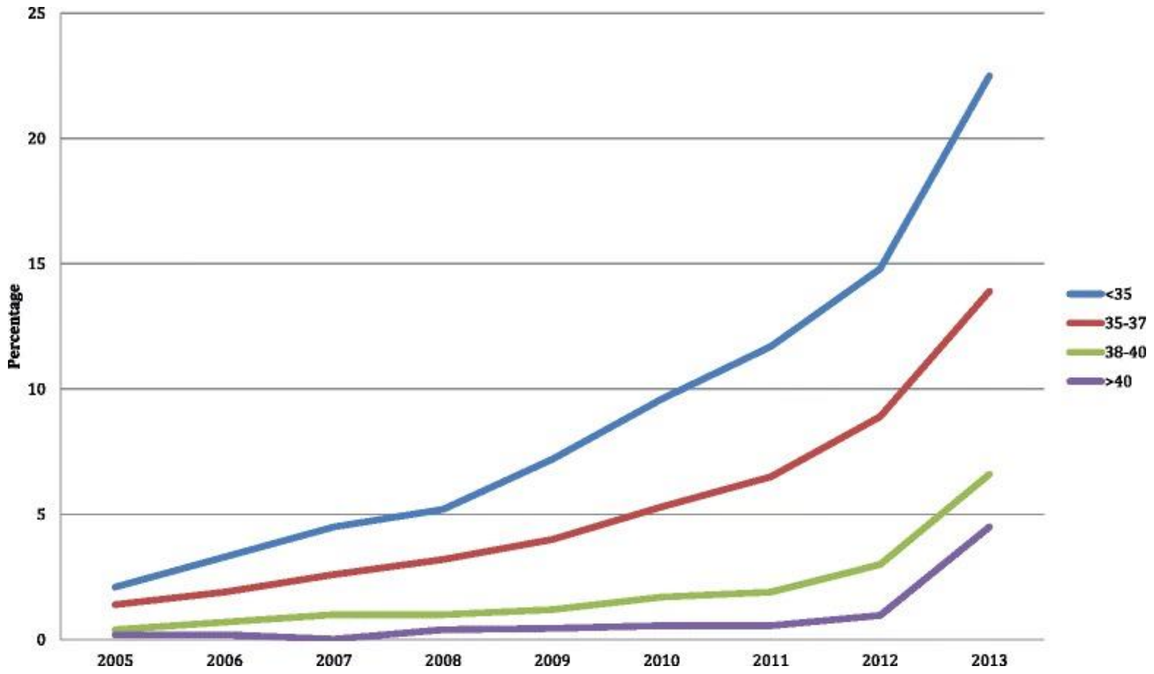


Figure 5: Proportion of IVF cycles with a single embryo in the USA from 2005-2013. Figure taken from Lee et al. (2016)[48], data from the Society for Assisted Reproductive Technology (SART) Clinic Outcomes Reporting System (CORS) database.

2.2 Embryo selection

There are a range of both invasive and non-invasive techniques available to assess the viability of embryos for transfer. Invasive techniques tend to have difficulties associated with them, for example pre-implantation genetic diagnosis and screening may have a negative impact on embryo viability[52]. Another potential embryo assessment technique is metabolomics, which involves examining the embryo culture media to detect the level of substances such as glucose, lactate or amino acids[53]. Moreover, these approaches are complex and time consuming so not practical for IVF clinics[54]. Therefore, the preferred procedure in IVF clinics is currently the non-invasive technique of time-lapse monitoring (TLM).

Time lapse Monitoring

There are two main benefits to TLM; 1) the embryos are less disturbed and 2) more information is available as they are recorded continuously. Daily embryo evaluations of morphology involve periodically removing the embryos from the optimal gas and temperature conditions, this is not necessary during TLM, and therefore embryos are under less environmental stress[54]. The increased temporal resolution gained by continually monitoring the embryos could in theory also be useful; for example the exact timings of the cell divisions can be assessed using TLM as the embryos are observed at every stage of their development so the key moments are not missed as they might be in standard procedure. We will first describe example TLM procedures and next discuss the experimental evidence forming the basis for these procedures.

Currently, there are three time-lapsing imaging systems available that are used in IVF clinics; EmbryoScope, Primo Vision and Eeva[55]. All three use a digital inverted microscope to record the development of the embryos, with photographs taken in intervals combined to create videos[55]. The EmbryoScope has a built in camera while Primo Vision and Eeva have a camera placed in the incubator.

The time-lapse videos produced are then used to assess the viability of the embryos. Many different grading systems have been devised to assign a score to each embryo, all of which use either or both the rate of development and the appearance of the embryos at specific stages to calculate the score[56]. One example proposed by Gardner et al. (2013)[56] that is based mainly on timing and the early morphology is shown in figure 6[56]. The embryos are sorted into classes A-E, or discarded immediately if the morphology is very bad. First, embryos that satisfied morphological exclusion criteria including first cleavage asymmetry, abrupt first division to more than two cells and multi-nucleation on day 2 of development were assigned to group E[56]. Any embryos not fulfilling the exclusion criteria were then sorted into classes A+-D depending on 3 timing variables; T3 - time taken for embryo to divide into 3 cells, T5 – time taken for the embryo to divide into 5 cells, and CC2- the duration of the second cell cycle.

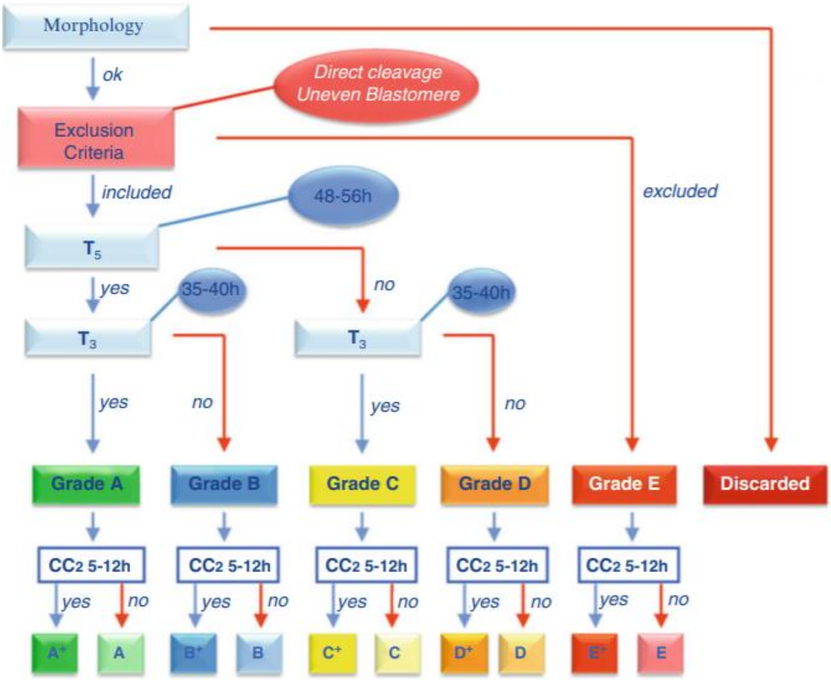


Figure 6: An embryos grading system based on timing and early morphology. T5 is time to 5 cell stage, T3 is time to 3 cell stage and CC2 is the duration of the second cell cycle. Taken from Gardner et al. (2013)[56]

Another example is the grading system devised by Gardner et al. (2000)[57] which is based solely on the appearance of the blastocyst. Each embryo is given 3 scores, a score from 1-6 depending on the degree of expansion of the blastocyst, a score from A-C for the ICM; A: tightly packed, many cells, B: loosely grouped, several cells or C: very few cells, and a score from A-C for the TE; A: many cells forming a cohesive epithelium, B: few cells forming a loose epithelium, or C: very few large cells.

A recent review, Armstrong et al. (2019)[58], of randomised controlled trials found that there is currently no evidence that the use of TLM gives a higher chance of a live birth than standard embryo incubation and selection procedures. However, this was mainly due to "high risk of bias in the included studies, imprecision, indirectness, and inconsistency"[58], as opposed to good-quality evidence disproving the benefit of TLM. Also, the potential detrimental effects of routinely removing the embryo from the incubator for imaging were not considered. There is a growing body of evidence to support the developmental origins of health and disease (DOHaD) hypothesis, which theorises that the environmental conditions during embryo/fetus development can have an impact on susceptibility to diseases as an adult[59]. Therefore, it is possible that the daily evaluations required without TLM, which place extra stress on the embryo, could have an adverse effect on the health of the child later in life.

Although there is no conclusive evidence supporting the use of TLM, many studies have been carried out using TLM to assess the impact of various morphokinetic parameters on different outcomes including blastocyst quality, implantation rates and live birth rates. As mentioned above, some of these have been incorporated into grading systems, although these tend to mainly focus on the appearance of the blastocyst. However, potential morphological indicators of viability have been reported at all stages of pre-implantation development; some of these will now be discussed in order of occurrence.

Zygote stage

Various studies have investigated links between morphological features at zygote stage and subsequent successful development of the embryo. One of the main points of focus at this stage has been the appearance and position of the PN[60]. Several studies have undertaken measurements to study the size of the PN and concluded that the overall size does not appear to have an impact on outcome[61], however the difference in size between PN has been shown to be important[62-65].

The appearance of the cytoplasm has also been studied; the cytoplasmic halo[60,66] and the presence of vacuoles[67] have been linked to embryo viability. Also of interest is the ZP, several studies have carried out measurements to evaluate the ZP thickness. The average thickness has not been reported to correlate to viability at this stage[61,68], however reports vary on the impact of thickness variation; greater variation in ZP thickness has been linked to higher chances of implantation[69] and clinical pregnancy[70]. However a study by Lewis et al[71]. found no correlation with ZP thickness variation and implantation.

Cleavage stage embryos

The timing of the first few divisions has been shown to be important in determining whether the embryo is able to properly form a blastocyst[72]. The symmetry of the divisions is also important[9], as if a cell splits unevenly one of the two cells produced by the division will have received less than half the cytoplasm, which can lead to a defective lineage in the embryo. It has been shown that 4- and 8-cell embryos with equal cell sizes have lower rates of chromosomal abnormalities and increased implantation rates[73-75].

Morula

The morula stage of pre-implantation development is understudied and often does not feature in embryo grading assessment procedures[76]. The presence of compaction at day three has been shown to be linked to embryo quality by a few studies[77-79], and the proportion of the embryonic matter included in the compacted morula and not lost to fragmentation has also been studied, with a higher proportion included linked to improved embryo viability[80,81]. The degree of compaction has not yet been linked to embryo quality, however this feature has been quantified by measuring the contact angles between blastomeres, with the contact angles shown to increase as the embryo undergoes compaction[82]

Another feature studied at this stage is the presence of vacuoles. A study by Ebner et al (2005)[67] found that vacuoles at morula stage were negatively correlated with blastocyst formation, and vacuoles had a bigger impact if they appeared in the morula than if they appeared earlier. Desai et al. (2000)[78] also included vacuoles at this stage as one of a few variables used to predict clinical pregnancy and implantation.

Blastocyst

As previously mentioned, the appearance of a single image of a blastocyst on day 5 is frequently used in IVF clinics, other features such as contraction rates may also be useful in assessing the viability of the embryo. The collapse of a blastocyst is caused by an efflux of blastocoel fluid due to loose cell bindings in the TE[83]. Marcos et al. (2015)[83] investigated whether such a collapse was related to viability, here a collapse was defined as a blastocyst contracting to such an extent that less than 50% of the surface of the TE was in contact with the ZP. They found that the implantation rates for embryos without collapse versus with collapse were 48.5% and 35% respectively. Vinals Gonzalez et al. (2018)[84] used the same definition of contraction to investigate the correlation between collapses and euploid embryos (healthy) vs aneuploid embryos (containing chromosome abnormalities). It was found that the total number of collapses was smaller in euploid embryos than aneuploid embryos (0.6 vs 1.57; $p < 0.001$).

In summary, there are many morphological features throughout pre-implantation development that may be useful in predicting transfer outcome. In table 1 we summarise all the markers we have discussed that have been shown to have a correlation with embryo viability.

Morphological feature	**Prediction**	**Reference**
Difference in PN size	Live birth	Otsuki, J. et al (2019)[62]
	Aneuploidy	Manor, D. et al (1999)[63]
		Sadowy, S. et al. (1998)[65]
	Implantation	Nagy, Z. P. et al. (2003)[64]
Presence of cytoplasmic halo	Blastocyst quality	Ebner, T. et al. (2003)[66]
Presence of vacuoles at zygote stage.	Blastocyst formation	Ebner, T. et al. (2005)[67]
ZP thickness variation at zygote stage	Implantation	Cohen, J. et al. (1988)[69]
	Implantation – no correlation found	Lewis, E. I. et al. (2017)[71]
	Clinical pregnancy	Sun, Y. P. et al. (2005)[70]
Timings of first few cell divisions	Blastocyst formation	Pera, R. A. R. et al. (2010)[72]
Evenness of blastomeres	Implantation	Van Royen et al. (2001)[74]
		Hardarson, T. et al (2001)[73]
Compaction at day 3	Implantation	Le Cruguel, S. et al. (2013)[77]
		Skiadas, C. C et al. (2006)[79]
	Pregnancy	Desai, N. N. et al. (2000)[78]
Degree of fragmentation at morula stage	Blastocyst formation	Ebner, T. et al. (2009)[80]
		Ivec, M. B. S et al. (2007)[81]
	Blastocyst formation	Ebner, T. et al. (2005)[67]

Presence of vacuoles at morula stage	Clinical pregnancy	Desai et al. (2000)[78]
Blastocyst collapse	Implantation	Marcos et al. (2015)[83]
	Aneuploidy	Vinals Gonzalez et al. (2018)[84]

Table 1: Summary of morphological markers that have been linked to embryo viability

2.3 Embryo abnormalities that can cause developmental failure

There are a wide range of developmental abnormalities that can cause either failure to implant or inability to continue to successfully develop post implantation. Two major cause of developmental failure are mutations of specific genes vital for development and whole chromosome abnormalities (aneuploidy).

Genetic mutations

As described in section 2.2, there are several transcription factors involved in the allocation of cells to each of the three separate lineages. Therefore, any genetic mutation that causes one of these transcription factors to be expressed differently or not at all could lead to developmental failure. One example that has recently been studied by Le Bin et al. (2014)[85] is the development of OCT4 knockout mice embryos. They showed that embryos where OCT4 had been deleted were still capable of forming an initial ICM, however the cells in the ICM were not able to further differentiate into either the PrE or EPI lineages, leading to eventual failure to develop.

Aneuploidy

Aneuploidy is the term used to describe an abnormal number or structure of chromosomes in a cell. A healthy cell should be euploid, which means the total number of chromosomes is an exact multiple of the number in a set[86] – for example 23 types of chromosomes in humans[87]. Aneuploidy occurs due to errors in the precise mechanisms involved in chromosome segregation during meiosis and mitosis. There are two main types; "whole chromosomal" aneuploidy, which is either a higher or lower than normal number of whole chromosomes, and "structural " aneuploidy – an abnormal structure of one or more chromosomes due to errors such as deletions or translocations of large regions of the genome[86].

Errors in chromosome segregation during meiosis are particularly high, especially in the production of oocytes. Around 10% of Oocytes are aneuploid and will therefore result in completely aneuploid embryos, most of which will fail to develop properly or result in an early miscarriage[87]. Whole chromosome aneuploidy of chromosomes 13,18 and 21 can result in live birth, however these will all carry birth defects – an abnormal number of chromosome 13 or 18 leads to severe developmental abnormalities and a life expectancy of less than a year and an abnormal number of chromosome 21 results in Down syndrome[86].

Aneuploidy in human embryos can also occur during the mitoses after fertilisation. This is thought to be the most common cause of aneuploidy[88] and may also directly causes early pregnancy failure[89]. It can lead to embryos with some aneuploid and some euploid cells, referred to as "euploid–aneuploid mosaicism"[89]. The proportion of aneuploidy cell has been observed to reduce throughout development, although it is not clear whether this is due to aneuploid cells selectively undergoing apoptosis, or embryos with a high proportion of aneuploid cells being more prone to developmental failure[89]. It has been shown that it is possible for mosaic embryos to result in a healthy live birth[90], however this has only been seen in a small number of cases as embryos known to be mosaic would not usually be transferred.

2.4 Other factors affecting success rate

The success rate of an infertile couple becoming pregnant varies widely between different clinics[91]. This is partly due to the different mix of patients between clinics, however this has been shown to not be the sole reason for the difference in success rates[91], suggesting that the chances of success pregnancy do depend on the methods and procedures of the clinics. There are also geographical differences, with higher success rates recorded in the USA than in Europe[92]. It has been suggested that this may be due to differences in the dosing of Gonadotropins[92], fertility medications given by injection to IVF patients.

The success of IVF also depends heavily on the age of the patient, pregnancy rates have been shown to go down steeply with maternal age while miscarriage rates rise[93] (figure 7). Paternal age has also been shown to have an effect on implantation and pregnancy rates, but is independent of miscarriage rate[94].

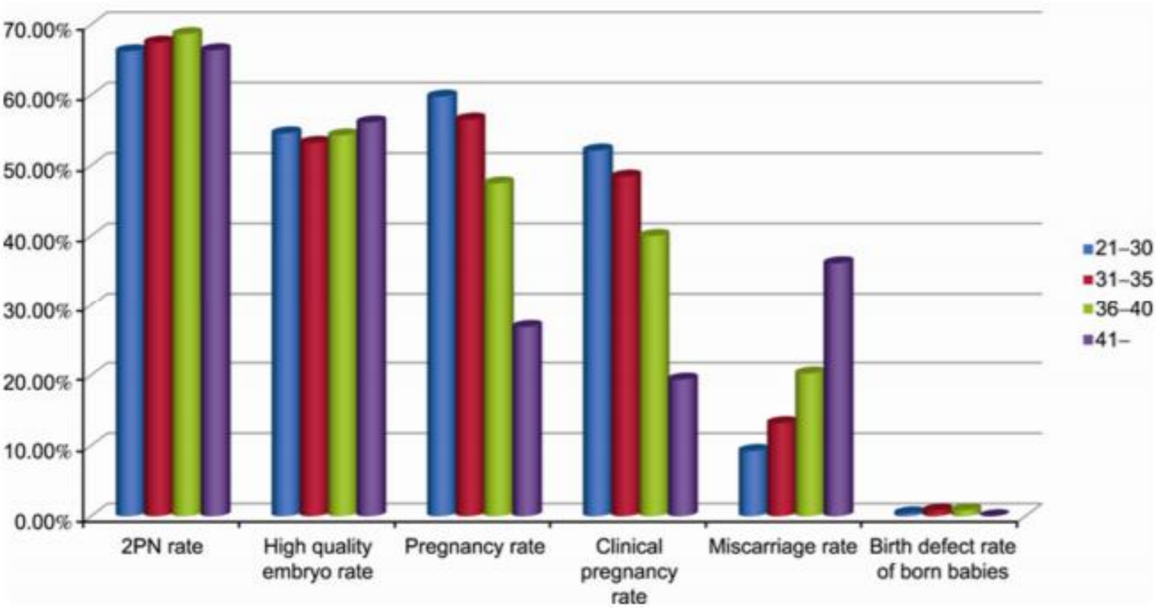


Figure 7: Effect of age on IVF outcome based on a retrospective study of 10268 women. The key shows age range of each age group From Yan et al. (2012)[93]

In addition to selecting the right embryo, gamete selection is also important. The quality of the selected oocyte has an impact on both successful fertilization and subsequent embryo development[95]. Sperm quality is also important, advanced sperm selection techniques have been shown to increase both fertilization and pregnancy rates[96].

3. Machine Learning

3.1 Overview

Machine learning is the science of computers being able to perform tasks without being explicitly programmed, a frequently used formal definition is "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."[97]. Machine learning algorithms can be broadly categorised into supervised, semi-supervised or unsupervised learning. In supervised learning, the training data consists of input – output pairs and the goal is to be able to predict the output for a given input[98]. In unsupervised learning, the data is unlabelled so rather than predicting an output the goal is to find patterns or structure in the data, often in the form of clusters[98]. Semi-supervised learning involves training an algorithm to make a prediction using both labelled and unlabelled data, usually used in cases where the amount of unlabelled data is much greater than the amount of labelled data[99].

Supervised Learning

There is a variety of supervised learning algorithms and the choice of algorithm will depend on the problem to be solved. The output can either be a number – known as a regression problem, or a class – known as a classification problem[100]. Regression problems can sometimes be solved with a linear function, where each input value is multiplied by a weight and a straight line is fitted to the data by iteratively finding the optimum value of the weights. If this model is too restrictive, a higher order polynomial or another non-linear function can be used[100]. However, as is shown in figure 8C, care must be taken not to over-fit the model to the data.

A popular classification algorithm is support vector machines (SVM)[101]. These work by drawing boundaries or “hyperplanes” between the classes that maximise the distance between the hyperplane and the datapoints in the class closest to the hyperplane, therefore minimizing classification error[101], as seen in figure 8A. Another popular classification algorithm is the decision tree, an example is shown in figure 8B. This sorts input data into classes depending on the value of the features of the data point. It consists of nodes, which represent the features to be classified, and branches which represent the different values that the node may take[101].

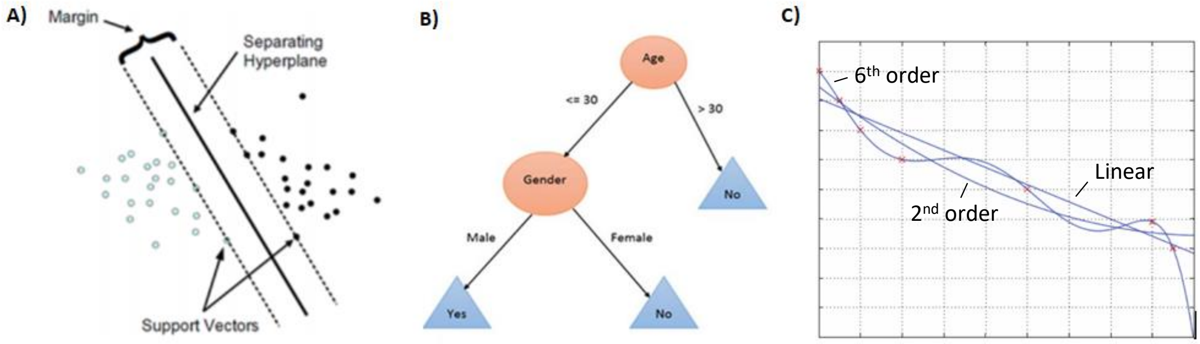


Figure 8: Illustration of machine learning techniques. A) shows support vector machine – an algorithm that maximises the difference between data-points and class boundaries, B) is a decision tree which finds the class of the data-point via a series of decision making "branches and C) is an example of regression using a linear, 2nd order and 6th order polynomial model. In this case the linear model is too simple and the 6th order polynomial over-fits the data, so a 2nd order polynomial is the best to use. A) and B) are taken from Dey (2016)[101], C) is taken from Alpaydin, Ethem. (2014)[100].

K-means

K-means is one of the simplest and most popular unsupervised learning algorithms that sorts data into a pre-determined number (*k*) of clusters[102]. The algorithm defines *k* centroids, one for each cluster, by using the following steps[102]:

1. Randomly assign *k* initial centroid positions.

2. Assign each data point to the group that has the closest centroid.

3. When all data points have been assigned, re-calculate the positions of each of the *k* centroids.

4. Repeat Step 2 and 3 until all centroids no longer move.

The goodness of the clusters can be evaluated by finding the average distance of each point to its assigned cluster centroid. The clustering algorithm will often be run many times from different initial centroids and the best attempt will be chosen as the algorithm can be prone to converging to local minimums.

3.2 Neural Networks

The neural network is a machine learning algorithm inspired by the biological concept of neurons[101]. A standard supervised neural network takes a list of input variables, runs these numbers through a series of functions, or "hidden layers" (see figure 9 for illustration), and produces one or more numbers as outputs[103]. Before entering each function, all inputs to the function will be multiplied by an individual 'weight'. These weights define how the inputs are combined and recombined[103], engineering new features which are fed into the final function to calculate the predicted output. The weights are initially set to random values, then are adjusted over many iterations using training data to minimise the difference between the output values and the predicted outputs. When the optimum weights are achieved the model can be tested on a validation training set to assess its accuracy before being used in real world applications.

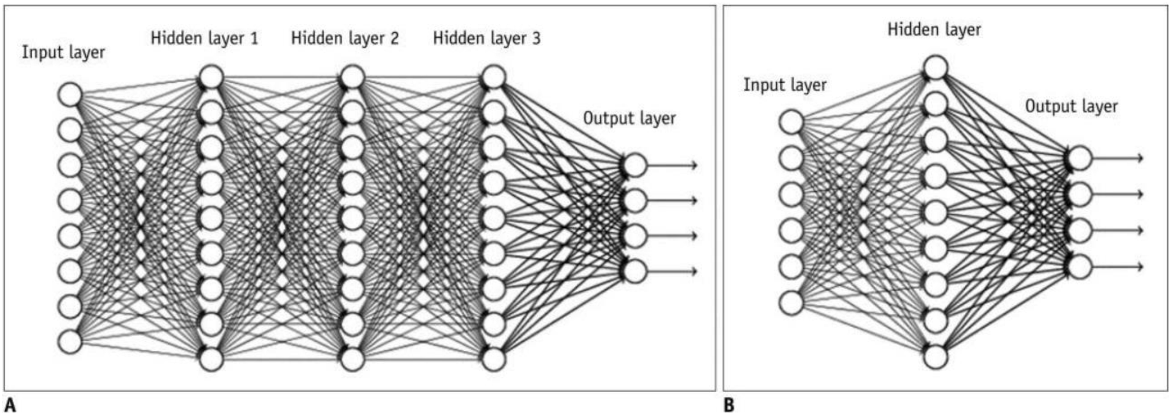


Figure 9: Two examples of neural networks. The nodes in the hidden layers all represent activation functions that take in inputs from every node of the previous layer and then deliver an output to every node in the next layer. Each input is multiplied by an individual weight before going into the activation function, the value of these weights are learned over many training iterations. A) is an example of a deep neural network (DNN), whilst B) is an example of a shallow neural network as it has just one hidden layer. Taken from Lee et al. (2017)[104]

There is a large amount of variety in the structure or architecture of a neural network, with different model architectures serving different needs. A neural network with many hidden layers is referred to as a deep neural network (DNN). These can be useful for extracting higher level abstractions from the data such as shapes or edges[103], which can be useful in complicated data sets that would have otherwise needed a feature engineering pre-processing step.

A convolutional neural network (CNN) is a specialised type of DNN that performs well for image recognition tasks[103]. CNNs are made up of a series of convolutional and pooling layers (see figure 10 for illustration) before reaching the final fully connected layer (those found in typical neural networks). In a convolutional layer one or more filters scan over the image and the weights in the filter are iteratively refined in the training process so that the filters can identify features in the input data. The pooling layers are used to reduce the dimensions of the outputs produced by a convolutional layer, either by taking the maximum or the average of a group of pixels.

Transfer learning

Transfer learning is a technique that allows information gained by training one CNN to be used in another one[105]. The process can typically be broken down into two[106]:

1) Train a model using a very large labelled (and usually pre-existing) dataset (referred to as a “base dataset”) to obtain a “base network” that can recognise a variety of images.

2) Train a new model using a smaller "target dataset" which contains only images of the objects that are required to be recognised. This model will inherit the first few layers of the base network so only the last few layers and any added layers will need to be retrained. The proportion of inherited, retrained and added layers will vary depending on the model and datasets.

The idea is that the first few layers of the base network will have been trained to recognise basic features that will be in all the target dataset images and the later layers will be able to use these features to detect more complex features specific to the objects in the target dataset and be able to then use these advanced features to distinguish between the different classes. As training the base network will often require a large number of images and can be very time consuming it is common to use pre-existing datasets and/or model architectures to train it. Keras, an open-source neural-network library, has at least 20 available models all of which have been trained on very large datasets[106]. This includes the VGG16 network which is very successful at image classification and has been trained on ImageNet, a visual dataset with over 14 million labelled images[107].

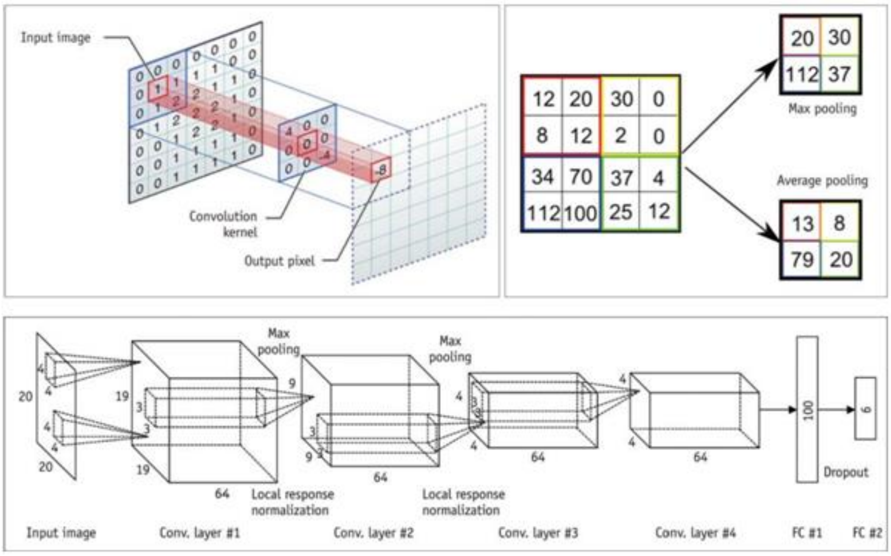


Figure 10: Illustrations of a CNN. The bottom picture shows an example structure of a CNN model, there are multiple convolutional layers interspersed with pooling layers to reduce the number of parameters. The top left picture illustrates a convolutional layer and the top right illustrates a pooling layer, showing examples of both max and average pooling. Taken from Lee et al. (2017)[104].

Transfer learning can be useful for various reasons[108]. One application is where the total data available for the target dataset is small, so the network is unlikely to be able to train itself correctly using just the target dataset. Another reason could be that it will be very time-consuming to label all the data-points in the target dataset, and a labelled base dataset is already available. Some researchers will also not have the computational resources or time to train a network from scratch, so using a pre-trained network such as the VGG16 network trained on ImageNet may be necessary even if a large labelled target dataset exists[106].

Interpreting Machine learning algorithms

It is very important that any machine learning model used to make decisions, especially important decisions such as medical diagnosis's, can be trusted to make correct predictions. All machine learning algorithms are tested on validation data sets; however it cannot always be assumed that a high accuracy on the validation set will translate to high accuracy in real life applications as there may be biases present. For example, there may have been information accidently included in the training and validation of the model that would not normally be available, known as "data-leakage"[109]. An example of this is the KDD-Cup 2008 breast cancer prediction competition, where the patient ID was shown to strongly correlate with both the training and validation data[109]. As the accuracy of validation sets cannot be completely trusted it is necessary to have some understanding of how the model makes predictions, which requires techniques for interpreting the model.

One such interpretation technique is LIME[110]. This technique approximates the machine learning algorithm locally with an interpretable model to explain a specific prediction. The interpretable model is found by minimizing the difference between the output of the original model and the output of the interpretable model, whilst still keeping the interpretable model sufficiently simple. A collection of sample points in the feature space are used, with those closer to the point in question being given a higher weight, this is illustrated in figure 11.

By looking at the interpretable model it is possible to see what input variables have the biggest effect on the predicted output, and the user can then use their own prior knowledge to decide whether this is sensible. An example is shown in figure 12, specific information about a patient is shown to be important for the model to diagnose the patient with flu, and the variables picked are all those that might be expected to be important.

Sometimes the original features are not used in the interpretable model; instead a different representation that is interpretable for humans is used. For images this may be the presence or absence of patches of similar pixels (super-pixels)[110]. A visual explanation of the model can then be given by showing an image that only contains the super-pixels that were important for a certain classification, with the rest of the image greyed out. An example of this is shown in figure 13, an image classified as "dog", "electric guitar and acoustic guitar shows the super-pixels you might expect for each classification.

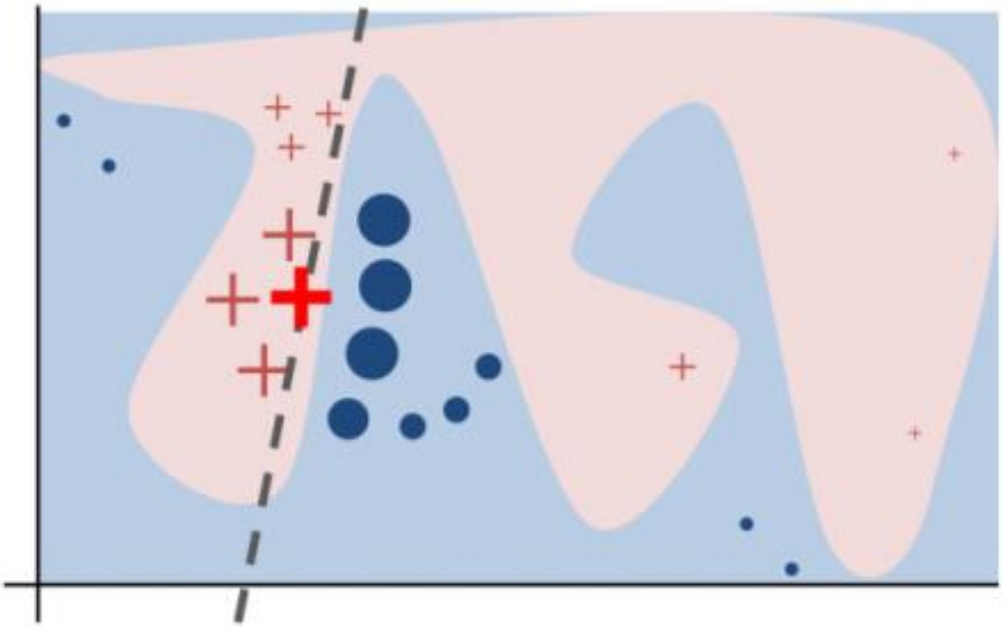


Figure 11: Illustration of an easily interpretable model found by LIME to approximate the original model. The blue and pink areas show the classification boundary of the original model while the dotted line is an interpretable model that approximates the original model in the local area shown by the red cross in bold. This interpretable model was calculated by sampling a selection of data points shown by the other crosses and dots. The sizes of each dot and cross represent how much weight they had in building the interpretable model, with bigger symbols having a greater weight due to being closer to the local area being modelled. Taken from Ribeiro et al. (2016)[110].

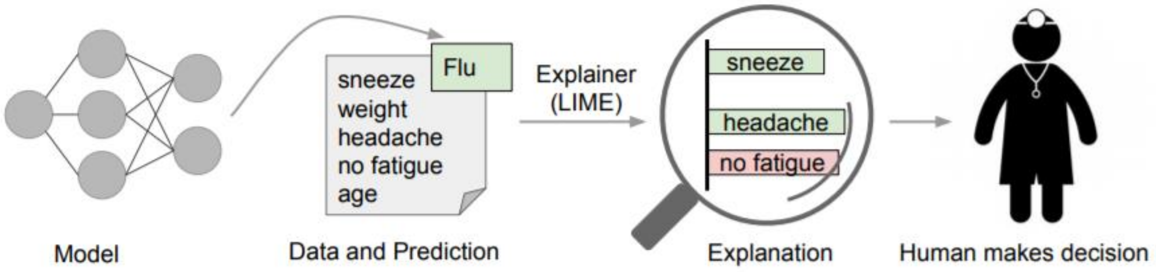
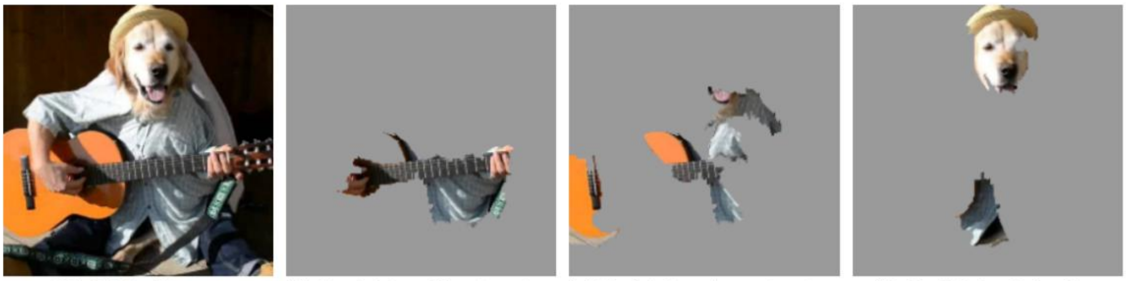


Figure 12: A conceptual understanding of how LIME gives an interpretation. The input variables that had the greatest effect on the "flu" classification are given in the explanation and suggest that the model is working properly as the variables highlighted seem to be sensible indicators of flu. Taken from Ribeiro et al. (2016)[110]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 13: An example of LIME explaining the predictions of a CNN. Images b)-d) show the parts of the image a) that were used to find their respective classifications. Taken from Ribeiro et al. (2016)[110]

3.3 Representations of image for machine learning – feature engineering

In order to minimise the processing time and computational costs of training a machine learning algorithm it is often necessary to first find a new representation of the data that contains the same information but using less features. This is especially important for deep learning algorithms which have a large number of parameters to train.

Principle component analysis

PCA is a linear technique that transforms the input data so that the many original variables can be represented by a smaller number of features[111] . This new representation with have the largest variances whilst still capturing most of the original information[112]. A very simple example is shown in figure 14; PCA is applied to 2D data so that instead of taking up two axes it can be described by just the x axis.

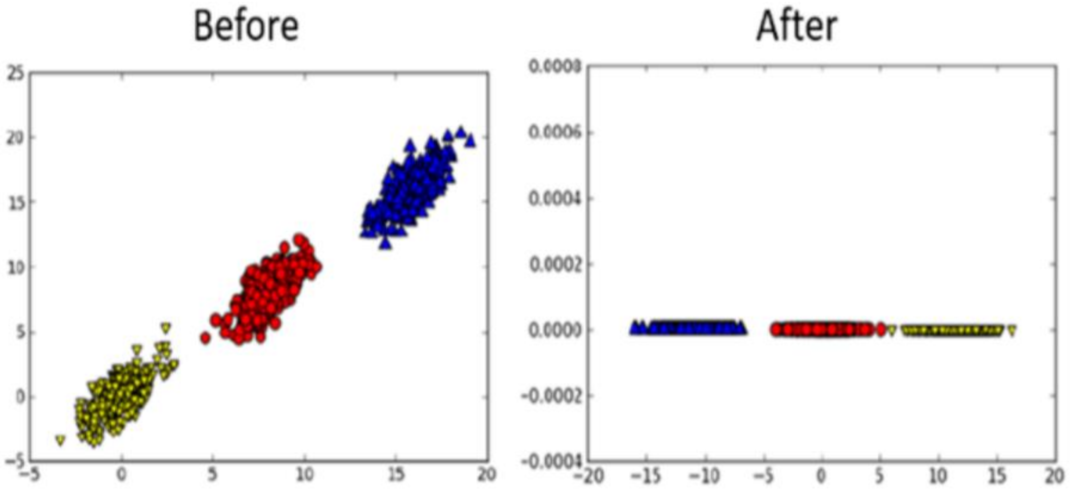


Figure 14: Visualisation of PCA applied to 2D data. Taken from Dey (2016)[101]

Autoencoders

An autoencoder is similar to PCA in that it produces a new representation of data that will be easier for a machine learning algorithm to process, however this transformation is done using an unsupervised form of neural network[113]. The inputs and the targets of an autoencoder are the same, so the job of the autoencoder is to reconstruct its own inputs[113]. To do this two parts are used; an encoder and a decoder[114], see figure 15 for illustration. The encoder finds the new representation and the decoder transforms the new representation back into the original inputs. The network is optimised by minimising the difference between the original inputs and the values returned by the decoder[114], referred to as the autoencoder loss.

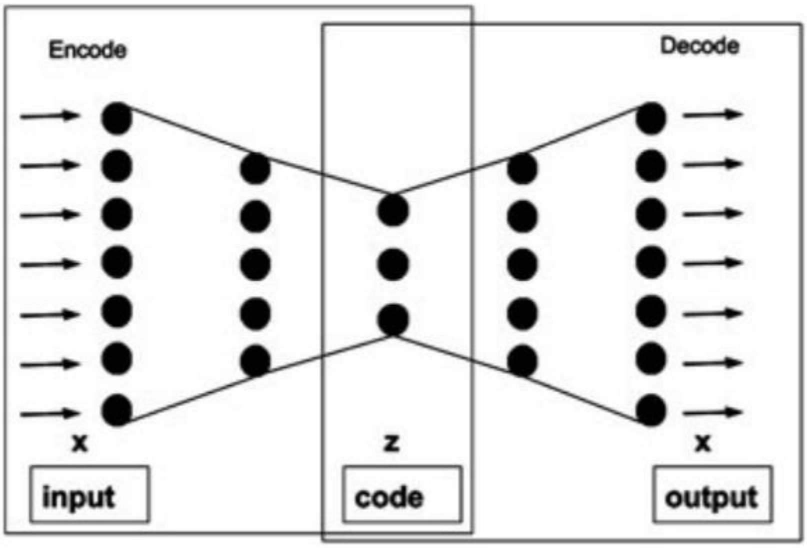


Figure 15: Visualisation of the structure of an autoencoder. The new representation is extracted from the middle layer (labelled z). Taken from Nasraoui et al. (2018)[115]

t-SNE

t-SNE is a dimensionality reduction technique that aims to preserve both the local and global structure of the data when reduced to 2 or 3 dimensions, mainly for visualisation purposes[116]. t-SNE is a variation of stochastic neighbour embedding (SNE)[117] that is easier to optimise and has less of a tendency to crowd points into the same space. SNE defines a probability distribution over all potential neighbours of an object by centring a Gaussian on each object. The aim is then to create a low dimensional space with a probability distribution as similar as possible to the probability distribution of the original high dimensional space by minimising the Kullback-Leibler divergence (A type of statistical distance measuring the difference between two probability distributions). t-SNE uses a symmetrized version of the SNE cost function with simpler gradients and uses a student t-distribution instead of a Gaussian for the lower dimensional space probability distribution[116]. t-SNE has a tuneable hyper-parameter, perplexity, which has a recommended value between 5 and 50 and is a rough estimate of the number of close neighbours each point has.

3.4 Machine learning and medical imaging

Since the 1980s a variety of machine learning algorithms have been developed to automate classification tasks in medical imaging and by the early 2000s many were introduced to clinical practice[104]. However, the benefit of using these tools was not clear due to issues found by clinical studies, such as the amount of false positives predicted by these systems[118]. It is hoped that deep learning methods may be able to overcome these issues and reliably predict diseases; deep learning tools have already been shown to have good potential to detect lung cancer[119], breast cancer[120,121] and Alzheimer's disease[122,123].

Although deep learning is not yet used in practice for embryo selection as part of IVF, there have been a few studies attempting to do this. The first attempt was made by Khosravi et al. (2019)[124], who trained a CNN to classify blastocyst-stage human embryos by quality. Time-lapse images of 10,148 embryos were used to train and validate the model and embryos were classed by embryologist assessment. They achieved an accuracy of 97.53% in differentiating between good quality and poor quality embryos. They also found that fair-quality embryos classified by the algorithm as high-quality had higher chance of live birth than fair-quality embryos classified as low quality (61.4% vs. 50.9%). However, attempts to directly predict the outcome using embryos labelled "positive live birth" or "negative live birth" were unsuccessful.

Since then many more studies have used deep learning to assess embryo viability, however most use intermediate outcomes such as embryologist assessment, aneuploidy, clinical pregnancy, foetal heartbeat, or implantation[3,125-135]. Only a few directly predict live birth[136-138]. Currently most deep learning studies also focus on the blastocyst stage, only a few use earlier developmental stages[128,131,136,139] and none predict live birth using only pre-blastocyst stage images.

4. Conclusions and Aims

In this chapter we have described current IVF practices and it is clear that the field is still developing, with one of the main limitations being the difficulty in selecting a healthy embryo for transfer. As mentioned in section 2.1, the preferred procedure now is to transfer just one embryo, making a robust selection procedure even more vital. Section 2.2 describes TLM, the current preferred method of assessing viability which has allowed many morphological features to be analysed as predictors of success rates. Although many possible links have been found and grading systems have been developed to assess embryo viability, there is still no consensus on embryo assessment and there is still much subjectivity involved. Therefore, machine learning may be the way forward; as described in section 3.4, machine learning algorithms have had success in a variety of medical imaging areas including some progress in embryo selection.

Our review of current knowledge of pre-implantation development shows that the mechanisms underpinning the formation of the first three lineages are still not properly understood. Further investigation into this period of development is needed to gain a greater insight into the events occurring. An improved understanding of these first few moments of development could help us in assessing embryo viability, and possibly lead to better diagnoses of the causes of infertility.

In light of the above, this PhD followed two main approaches; developing ML algorithms that could predict transfer outcome of IVF embryos and furthering current knowledge of mammalian pre-implantation development. In chapter 2, CNN models are developed to predict live birth from a variety of embryonic stages from zygote to blastocyst, with the optimal moments in development for embryo assessment identified. Chapter 3 builds on these results by investigating a variety of morphological features that may be predictive of embryo viability and combining these with the CNN predictions to obtain a further improved prediction of live birth at early embryonic stages. Chapter 4 then investigates variations in the appearance of the nucleus, a currently neglected morphological feature that might also be incorporated in live birth prediction in the future. For chapters 2 and 3 time-lapse videos were provided by St Marys IVF clinic, Manchester, and for chapter 4 we produced new images of mouse embryos. Finally, we reflect on the research carried out in this thesis in chapter 7, discussing the relevance of our findings and potential avenues for further work.

The overall hypothesis of this thesis is that machine learning techniques can lead to improvements in the embryo selection process. The specific aims are; 1) Develop a machine learning model that can predict live birth from time-lapse images, 2) Assess the possibility and potential benefits of predicting live birth at pre-blastocyst developmental stages, 3) Investigate additional morphological features that can be combined via machine learning to give better assessments of embryo viability, and 4) Establish typical nuclei appearance over development and cell lineage in mouse embryos to pave the way for possible future inclusion of nuclei appearance in embryo viability assessments.

Chapter 2:

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

Camilla Mapstone, Helen Hunter, Daniel Brison, Julia Handl, Berenika Plusa

Rationale of paper

To address the issue of subjectivity in embryo selection for IVF we developed CNN models to predict whether embryos would lead to live birth after transfer based on images extracted from time-lapse videos. As earlier stages have so far been neglected in ML studies for embryo selection, we developed models at various stages in development, both to provide the option of earlier selection and to ascertain whether better quality assessments could be made at blastocyst stage when information from earlier stages is taken into account.

Another motivation of this paper was to show that it is possible to develop DL models for embryo selection from a relatively small single-clinic dataset. This would allow models to be re-trained to become tailored to specific patient populations. Our blastocyst model achieved a similar performance to highly trained embryologists, demonstrating that it is indeed possible to successfully train a CNN for live birth prediction on a dataset of this size.

This manuscript has recently been submitted to npj Digital Medicine and has been sent to reviewers. All supplementary figures and tables are included at the end of the manuscript.

Aims

-Develop CNN models that can predict live birth from a variety of individual stages from zygote to blastocyst

-Investigate whether models trained at pre-blastocyst stage offer any potential improvements to blastocyst stage selection

-Demonstrate that a CNN can be trained to predict live birth at a high standard from a single-clinic dataset

Author contributions

Dr Berenika Plusa, Prof. Julia Handl and Prof. Daniel Brison contributed to the concept of the study. Helen Hunter retrieved the time-lapse data from the IVF clinic. I pre-processed the data, developed the models and analysed the results. Dr Berenika Plusa, Prof. Julia Handl and myself designed the eMLife pipeline. I drafted the manuscript and then subsequently edited it before submission to address feedback provided by Dr Berenika Plusa, Prof. Julia Handl, Prof Daniel Brison, and Helen Hunter.

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

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

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

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

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

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

Abstract

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

Main Text

Infertility is a growing health crisis impacting both individuals and society. As a result, there is a rising demand for in vitro fertilisation (IVF) treatments. However, due to low success rates multiple attempts are often required, leading to additional cost and distress for the patients. One of the main challenges in IVF is the difficulty in selecting the best embryo to be transferred, as at the moment there is no general consensus on exactly what healthy pre-implantation embryo development looks like.[1-3] The selection process routine in most clinics involves visual assessment of embryos in real time or via time-lapse videos. Embryos are assessed based on morphological features such as the number and size of cells at cleavage stage and in the trophectoderm (TE) and inner cell mass (ICM) in the blastocyst (Fig. 1A), the expansion of the cavity, developmental timings, cellular fragmentation, and multi-nucleation. This manual assessment is subjective, with up to 83% variation between embryologists[4], and uses only a fraction of the information potentially available. A machine learning (ML) approach that automatically assesses embryos using more extensive information from across the time-lapse videos could potentially provide a consistent and reproducible method for embryo selection[5]. ML has already demonstrated great promise in other areas of medical imaging[6-8] and has the potential to revolutionise the field of medical diagnostics. However, existing ML approaches for embryo selection[5] still fall short in successful prediction of live birth, the lack of attention to pre-blastocyst stages, and the need for reproducible models trained on single clinic data.

Most previous attempts to train ML models to assess embryos from raw time-lapse data aim to classify embryos based on intermediate outcomes such as manual grading by embryologists or embryo aneuploidy, or early pregnancy outcomes after embryo transfer such as foetal heartbeat[4,9-21]. Only a few have been designed with the outcome of live birth[22-24]. Predicting live birth directly is a difficult task as there are a large variety of factors that contribute to treatment success, some related much more strongly to maternal factors than the embryo itself[25]. However, it is important to predict live birth as it is the overall aim of IVF and intermediate outcomes do not guarantee success. For example, it possible for embryos with good morphology to be aneuploid[26,27], the presence of some aneuploid cells in an otherwise diploid embryo (termed mosaicism) does not guarantee embryo failure[28], and even if the transferred embryos progress to foetal heartbeat, they may be miscarried later in pregnancy.

Currently most ML studies solely focus on the blastocyst stage, with very few using earlier embryonic stages[11,14,17,24] and none making a live birth prediction on embryos before the blastocyst stage, mirroring the most widely used process of embryo selection by the trained embryologists. Extending the culture until blastocyst stage is associated with a significantly higher live birth rate, per transfer and per embryo[29], and yet blastocyst transfer has a number of disadvantages. The first is that the cumulative live birth rate, including transfer of all fresh and frozen embryos from a single egg collection event, have not increased using this policy, due to the loss of embryos during extended embryo culture[29,30]. The second and more problematic issue is that blastocyst culture is associated with a number of adverse outcomes from pregnancy and early childhood, including pre-term birth, altered birthweight, and mono-zygotic twinning[30-33]. It is also possible that blastocyst culture may have negative implications on healthy aging, as recently it has been linked to shortened telomeres[34]. Adopting ML-assisted protocols that would allow for embryo selection at the earlier stages could be a step forward to reduce the risks that accompany extended embryo culture in IVF procedures.

A significant barrier to the widespread adoption of ML methods in ART is the depth and breadth of data required to develop an algorithm sufficiently predictive to be of use clinically. Many previous attempts have used large aggregated datasets derived from a number of clinics over a number of years in order to generate sufficient data to train ML models[11,19,20]. This in turn is a major limitation to widespread adoption of ML in the sector, as most clinics have distinct treatment policies and patient populations[35] . There is a need for a versatile and robust ML algorithm for embryo selection that can be successfully tailored to the needs of each individual clinic.

Here, we present eMLife, a novel pipeline for creating a versatile live birth predictor tool. Firstly, through the development of a stage classification algorithm we demonstrated the high efficiency of our chosen model at correctly classifying stages of pre-implantation embryos. Next, we used this ML algorithm to identify pivotal time-points for live birth prediction in the pre-blastocyst embryo, paving the way for earlier embryo assessment and transfer. We then showed that our ML based pipeline can assess blastocyst viability at a similar level to highly trained embryologists. Finally, we showed that combining information from our identified time-points of embryo development with blastocyst stage predictions allows for a quantitative ranking of high-quality blastocysts which could not have been achieved by any of the existing assessment methods. Critical to widespread clinical application, our findings also demonstrate that it is possible to train a ML model to a high standard using very limited data from a single clinic, paving the way for clinic specific, ML-assisted assessment of embryo quality.

Results

All model training and testing was performed on a dataset of time-lapse videos of embryos with known transfer outcomes from the IVF clinic at the Department of Reproductive Medicine, St Marys' hospital, Manchester, UK. This is an NHS-funded clinic with stable patient population demographics and standardised treatment policies, and therefore highly suitable for single centre treatment outcome studies[36]. Embryos were cultured in either Embryoscope ™ or Embryoscope+ ™ time-lapse incubator system (Vitrolife, Sweden).

Developmental stage classification

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

To develop stage classification models, all time-lapse videos had one frame extracted for each developmental stage, resulting in equal class sizes. To reduce the amount of training data needed, we adopted a transfer learning approach, using the MobileNetV2 model[37] (chosen following a preliminary analysis into the performances of several different models on our dataset) with layers pre-trained on ImageNet, the overall workflow is illustrated in Fig. 1B.

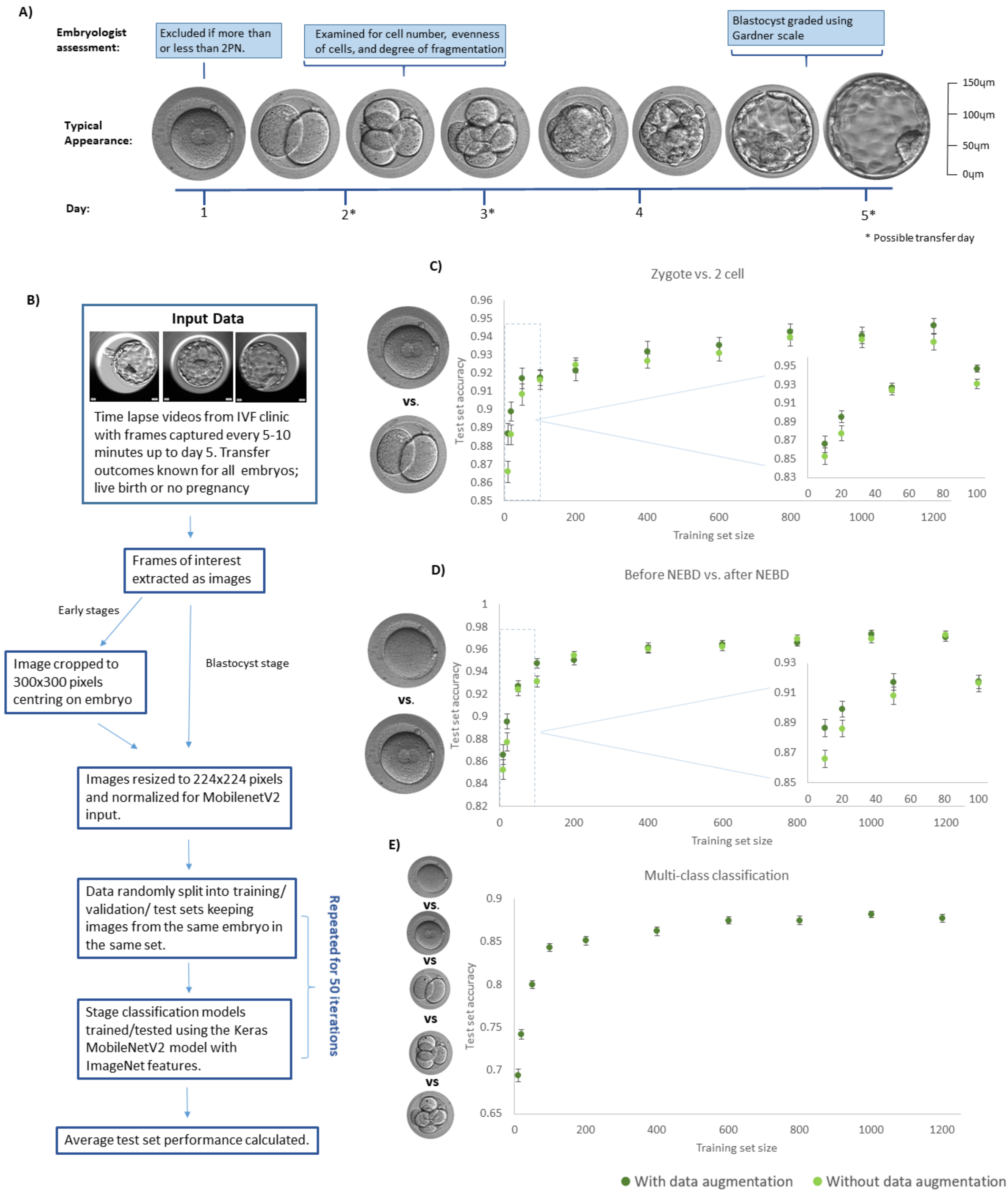


Figure 1: Performance of developmental stage classification models with increasing amount of training data A) The typical development of human embryo and an overview of embryologist assessments carried out at each stage. B) The methodology of training developmental stage classification models from time-lapse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

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

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

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

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

We then trained multi-class models to classify images into five output classes corresponding to the consecutive stages of human pre-implantation development (Fig. 1E). We saw that an accuracy of 87.7% was achieved, which is much higher than the by-chance score of 20%. The results in Fig. 1E show a similar trend to the one observed in the binary models; increases in accuracy became small once the training set reached about 200-400 images, and a high accuracy (69.4%) was achieved with a training set of just 10. These results further demonstrate that MobileNetV2 with pre-trained layers can achieve a high performance, even with a small amount of the training data, when analysing routinely collected images of preimplantation human embryos.

Live birth prediction

Following the successful development of our stage classification models, we trained models to predict live birth outcome (Fig. 2A). Models were trained on a variety of both blastocyst and pre-blastocyst (zygote, cleavage and morula) stages as this could allow for information from across development to be included in embryo assessment and also could potentially assist earlier selection and transfer in the future.

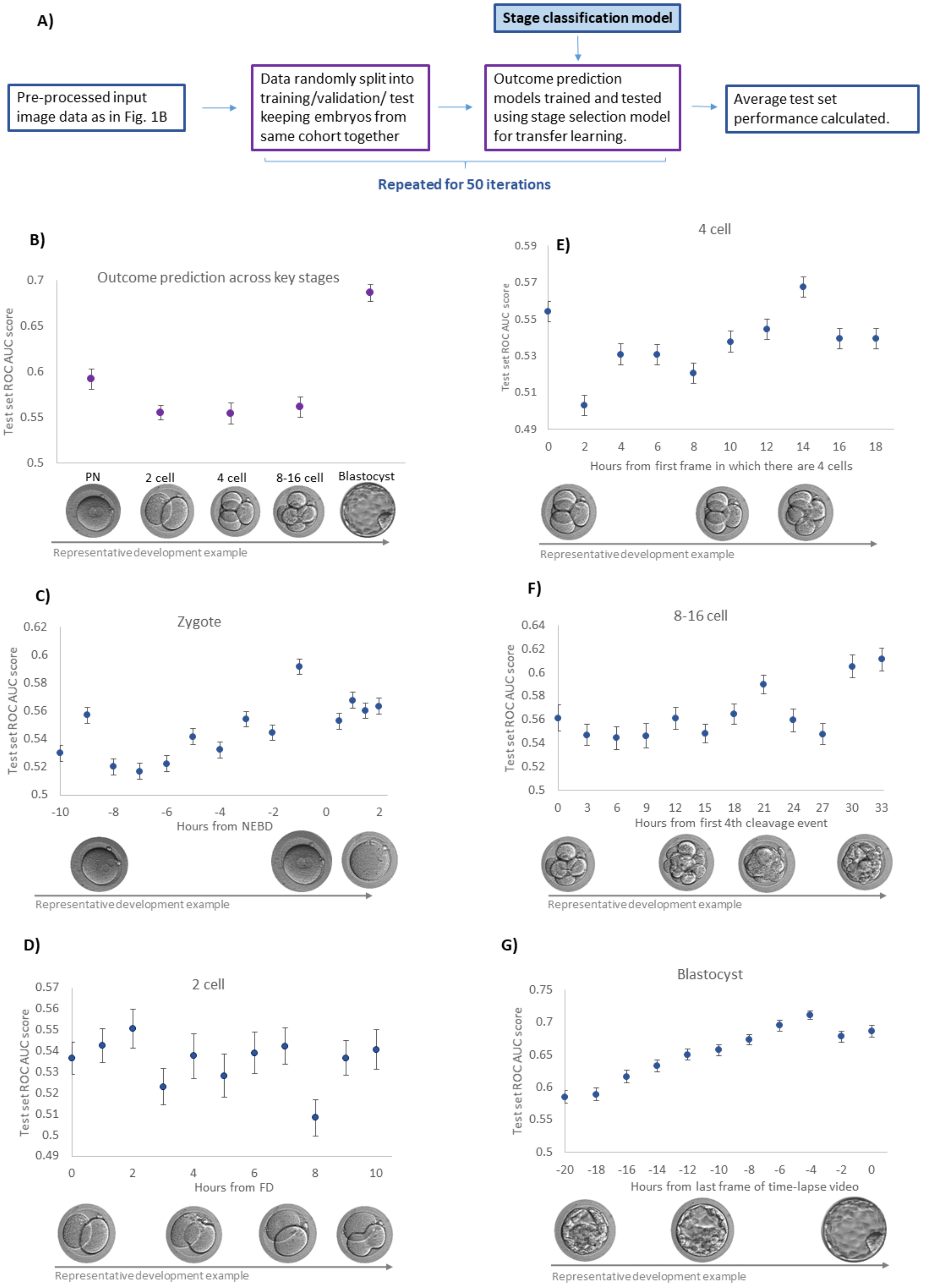
We extracted frames at five precisely defined stages; one hour before NEBD completion, first appearance of 2 cells, first appearance of 4 cells, initiation of 8-16 cell division round, and the last frame before the embryo was removed for transfer, and we referred to these stages as PN, 2-cell, 4-cell, 8-16 cell, and blastocyst respectively. Live birth prediction models were trained using each stage. As predicting live birth from embryo morphology alone is a difficult task, we decided to use our multi-class stage classification model as an extra transfer learning step (see methods section for full details). We experimented with various hyperparameters, the results are shown in Fig. S2A. Our results suggest that this extra transfer learning step generally improved model performance, particularly for the PN and 4-cell models.

We assessed our models using ROC AUC values; the area under a curve (called the ROC curve) that is created by plotting true positive rate vs. false positive rate at various thresholds. A ROC AUC of 0.5 is no better than chance and 1 is a perfect model. The ROC AUC values reported in Fig. S2A were the average of 50 training iterations each with a different randomly selected test set. We chose to use varying test sets to increase statistical power as we found that separate test sets had a higher than desirable variability in performance, reflecting the heterogeneity of the data in this ML task (Fig. S2b). However, the transfer model was trained on the same dataset, so to check this was not resulting in an unfair bias it was necessary to also develop all models from scratch with a true 'hold out test set' – a test set separated out at the very start, prior to training the transfer model. As shown in table 1, our results continue to hold up in this control experiment, so going forward we use transfer learning (with our optimised hyper-parameters) for the PN and 4-cell models and the original MobilenetV2 model for all other stages.

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

Table 1: Effect of extra transfer learning on hold-out test set. Performance of a hold out test set using either the original MobileNetv2 model with fixed features up to the last layer or the MobileNetv2 model with fixed features and

extra transfer learning; an extra hidden layer before the last layer has been pre-trained as a stage classifier. The number of hidden units in the last layer is the amount we found to be optimal for that developmental stage.



Pre-processed input image data as in Fig. 1B

Data randomly split into training/validation/ test keeping embryos from same cohort together

Stage classification model

Outcome prediction models trained and tested using stage selection model for transfer learning.

Average test set performance calculated.

Repeated for 50 iterations

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

The average test set ROC AUC over 50 training attempts with random train/validation/test split is shown in Fig. 2B for each stage. The blastocyst stage has the most obvious visual difference between classes, with unsuccessful embryos less likely to form expanded blastocysts, and as expected this model had the best performance. However, the pre-blastocyst models still gave above chance predictions. From our model performance it was clear that different developmental stages carry different predictive power. To identify specific moments in development for which life birth predictions are the most successful we tested multiple frames at regular intervals using each of the previous developmental stages as reference points (Fig. 2C-G).

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

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

Table 2: Percent of embryos that have begun cavitating. Shown for each outcome group at various time-points after first 4th cleavage event

Finally, to test the potential benefit of averaging over repeat training attempts and also obtain a single prediction for each embryo in our dataset (to be used in further investigations reported below), we calculated an average score for each embryo using each model. The PN, 2-cell and blastocyst stage models were trained using the original time-point and the 4-cell and 8-16 cell stage models were trained using the peak performance time-point (4-cell +14hr and 8-16 cell +21hr). To do this we retrained each model five times (with 50 repeat training runs) using five mutually exclusive test sets that spanned the whole dataset when combined. For each embryo, the 50 resulting model confidence scores from the 50 training iterations were averaged to produce a final 'average test set score' for that embryo. ROC AUC values were then calculated for each model using these average test set scores. This showed (table 3) that slightly better predictions can be made when the confidence score assigned to an embryo is averaged over many separately trained models, which is consistent with the existing literature on ensemble learning[38].

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

Table 3: Effect of averaging score from many models

Comparison of ML performance to Embryologist selection

To compare our results against embryologist selection performance we used a subset of embryos for which Gardner grades[39] were available (141 in total) to produce ROC curves using blastocyst model average test set scores and embryologist grades. In order to produce the ROC curve from embryologist grades we first converted the Gardner score letter grades assigned to the ICM and TE into numbers, as proposed by Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology[40]. Then an overall score was obtained by calculating an average of TE, ICM and expansion scores.

The ROC curves for both blastocyst assessment methods are shown in Fig. 3. The ROC AUC was similar for embryologist and model assessment, 0.720 and 0,726 respectively. This suggests our model had a very

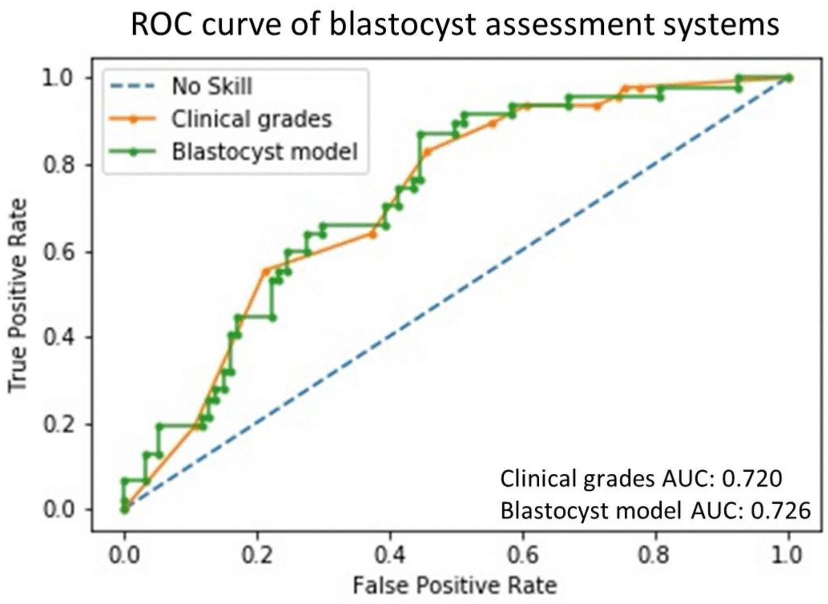


Figure 3: The live birth prediction ROC curves and AUC values obtained using the clinical grading system and the blastocyst model. Only embryos for which clinical grades were available were included, a total of 141. Clinical grades were originally based on the Gardner scoring system[39], these were then converted to numerical grades to allow a ROC curve to be generated. The blastocyst model predicts live birth using the last frame of the time-lapse video on day 5, the predictions produced by this model are a number between 0 and 1, 0 corresponding to an unsuccessful transfer (no pregnancy) and 1 corresponding to a successful transfer (resulting in live birth).

similar performance to highly trained embryologist grading, as St Marys' clinic host the UK NEQAS (a charitable consortium of external quality assessment laboratories) in reproductive biology. However, the blastocyst ML model was at a disadvantage due to the use of limited information; only a single focal plane from just the final frame was used to train the ML model. Whereas the assessment of the embryologist was done using multiple focal points from the final frame, with possible adjustment of the score according to the time-kinetic data.

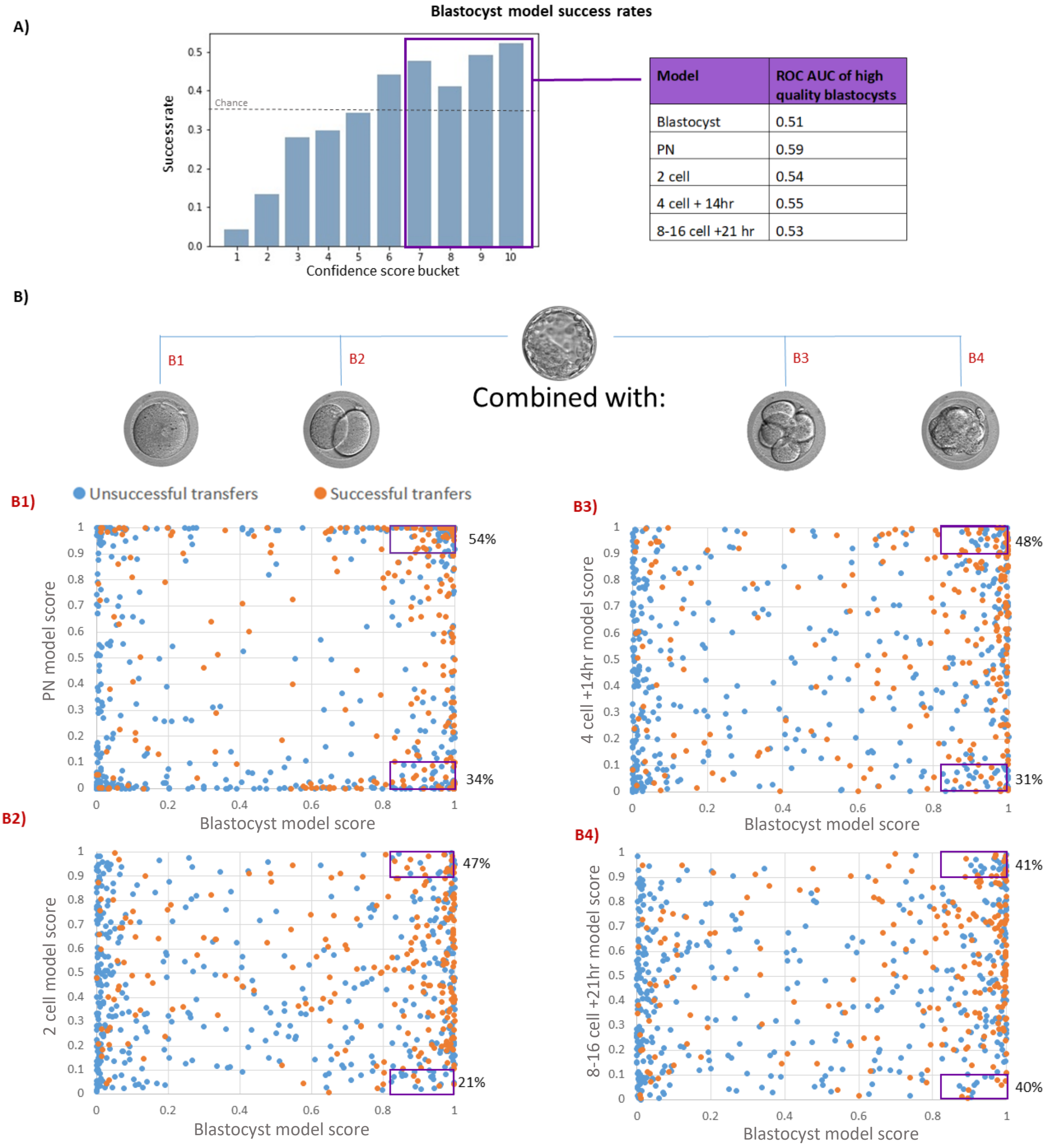
To investigate the effect of using just one focal frame for each embryo, we examined the confidence scores given to embryos where the ICM was not visible but the TE appeared to be good quality (Fig. S3A). We found that these embryos tended to receive very high confidence scores, suggesting that the ML model may not have used the ICM to make predictions. We then tested this further using the LIME software[41], which illustrates which parts of an image most strongly influenced a model classifying that image. The explanation images produced (Fig. S3A) appeared to focus on the TE, possibly using the size of the cavity to determine blastocyst quality. The ICM was often not included in the areas highlighted as useful to the model at all, providing further evidence that the model may not have been using ICM data.

Combined model outputs allow refined ranking of high grade blastocysts

In clinical practice, a numerical scoring system of embryo viability is more useful than a binary 'successful' or 'unsuccessful' prediction, as it allows the embryos in a cohort to be ranked in order of likelihood of live birth. Although our models were trained as binary classifiers, the confidence score is continuous so potentially could be used to rank embryos. We investigated this for our best performing model, the blastocyst, by sorting all the embryos in our dataset into buckets by blastocyst model average test set score, bucket 1 being the lowest score and bucket 10 being the highest score. The success rate (number of successful embryos in bucket/total number of embryos in bucket) for each bucket is shown in Fig. 4A. There was a general increase in success rate from bucket 1 to 10, suggesting that the confidence score was correlated to chance of successful transfer and therefore could be useful as a method to rank embryos in a cohort.

Despite a positive correlation, Fig. 4A showed that the blastocyst score was still not a perfect indicator of success, particularly when the score was high. Therefore, we then investigated whether the confidence scores from models trained on the earlier stages could add extra information to help chose between embryos with a similar blastocyst score. As increase in success rate with bucket number appeared to plateau from bucket 7-10 we hypothesised that pre-blastocyst ML model scores might be more useful than the blastocyst model to choose between embryos falling in this window (corresponding to a blastocyst model score of >0.83). To test this, we calculated the ROC AUC using models at each stage on this subset of 'high quality' embryos. The results (also shown in Fig. 4A) show that the blastocyst model performed almost no better than chance and worse than all the earlier stage models, and the PN model performed the best. This suggests that if multiple embryos in a cohort fall into this high quality blastocyst sub-group (blastocyst model score>0.83), then pre-blastocyst stage models should be used to choose which one of these high quality blastocysts to transfer, rather than simply choosing the one with the highest blastocyst score.

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



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

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

Discussion

IVF success rates are currently restricted by the difficulty in selecting the best embryo for transfer, and developing ML algorithms to assist with this task is an active area of research[5]. The ML pipeline presented here, eMLife, is the first to use early stages of pre-implantation human development prior to the blastocyst to predict live birth outcome following transfer, and is also the first to demonstrate these pre-blastocyst predictions can be used to refine the selection for transfer of high quality blastocysts. eMLife is reproducible and can inform multiple transfer strategies, therefore could lead to a universally applied standard protocol for embryo selection (Fig. 5). Our work has made contributions to three important areas; predicting live birth directly, assessing the predictive potential of pre-blastocyst embryonic stages, and developing a ML algorithm with a small routinely collected dataset from a single clinic over a restricted period of time.

It is important to note that all embryos included in the dataset used for this study were created by ICSI. To ascertain whether the models trained in this study can give a similar performance for embryos created from traditional IVF it would be necessary to validate the models on a dataset including IVF embryos.

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

There have been many ML studies assessing embryo quality in recent years, however only a few have attempted to directly predict live birth (using a dataset exclusively containing transferred embryos)[22-24]. Two studies have attempted to predict live birth using images extracted from timelapse at various time-points; Ueno et al.[23] and Sawada et al.[24]. Ueno et al. used the deep learning model IVY, which had been developed by Tran et al.[11] to predict fetal heartbeat using 10638 embryos. Using IDAscore, an algorithm based on IVY, Ueno et al. attempted to predict live birth and reported a ROC AUC varying from 0.66 for age 41-42 to 0.76 for age>42. No average score for all ages was given so it is not possible to give a direct comparison, however our highest performing model; - the blastocyst model, achieved within this range (0.69). Sawada et al. achieved a ROC AUC of 0.64, lower than our blastocyst model, however it should be noted they had a very limited amount of data available (only 91 successful embryos). Miyagi et al.[22] predict live birth from just blastocyst images, so is the most directly comparable to our blastocyst model. However, they also only report ROC AUC by age, with the value ranging from 0.634 for under 35 to 0.866 for over 42. No overall result is given, but again our blastocyst model falls within this range.

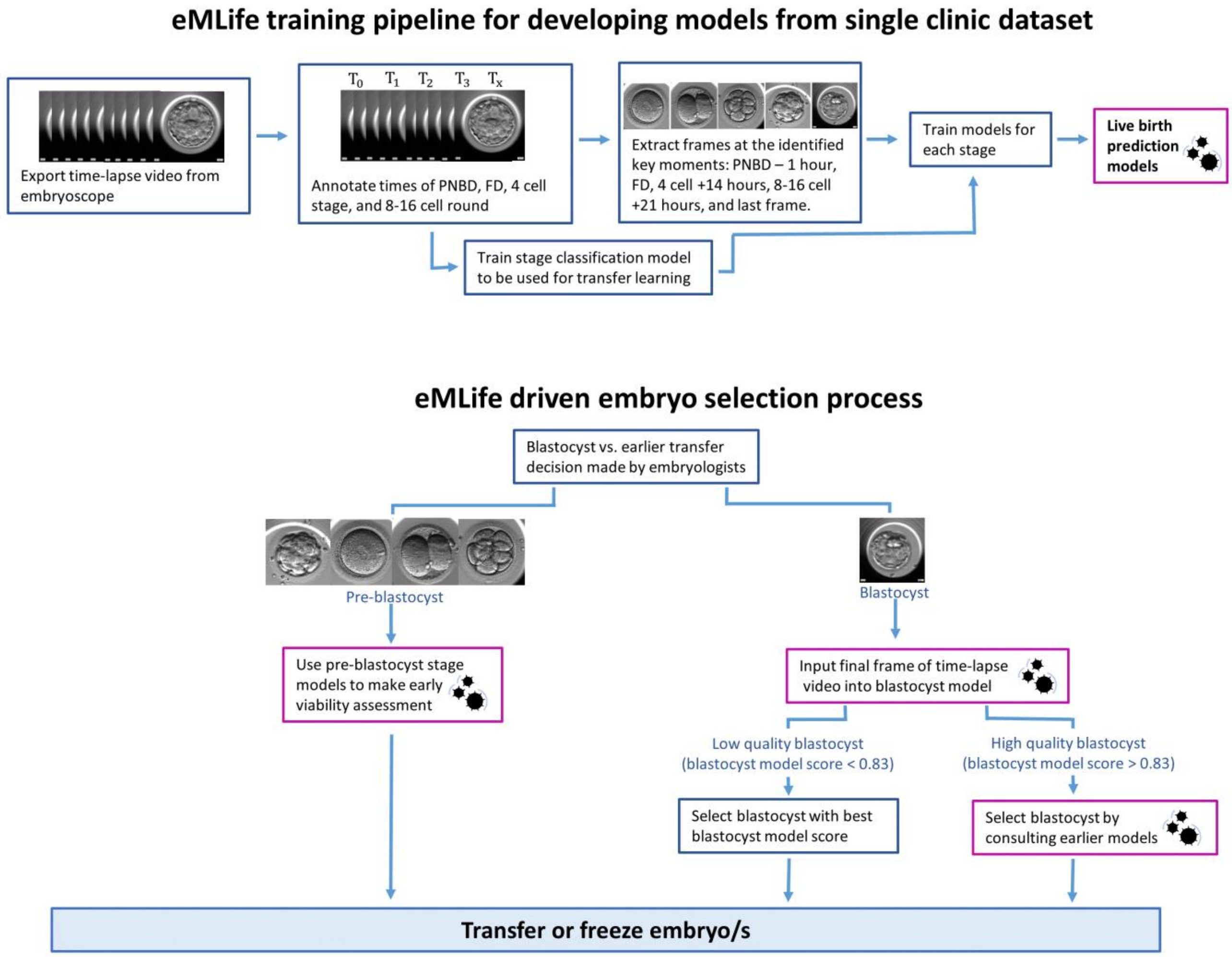

eMLife training pipeline for developing models from single clinic dataset
Export time-lapse video from embryoscope
T_0 T_1 T_2 T_3 T_x
Annotate times of PNBD, FD, 4 cell stage, and 8-16 cell round
Extract frames at the identified key moments: PNBD – 1 hour, FD, 4 cell +14 hours, 8-16 cell +21 hours, and last frame.
Train models for each stage
Live birth prediction models
Train stage classification model to be used for transfer learning
eMLife driven embryo selection process
Blastocyst vs. earlier transfer decision made by embryologists
Pre-blastocyst
Blastocyst
Use pre-blastocyst stage models to make early viability assessment
Input final frame of time-lapse video into blastocyst model
Low quality blastocyst (blastocyst model score < 0.83)
High quality blastocyst (blastocyst model score > 0.83)
Select blastocyst with best blastocyst model score
Select blastocyst by consulting earlier models
Transfer or freeze embryo/s

Figure 5: Training and deployment of the eMLife pipeline Top: eMLife workflow for training live birth prediction models for various developmental stages. Bottom: Proposed new approach to embryo selection using the eMLife models.

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

Currently, most embryo assessment approaches are focused on the blastocyst stage, especially when using ML algorithms. Here, we have shown it is possible to predict live birth from stages earlier than the blastocyst, with ROC AUC values of 0.578-0.603. These pre-blastocyst predictions could be useful for various transfer strategies; they could either be used to improve selection at blastocyst stage (as was successfully demonstrated in this work), or they could assist selection at an earlier stage, allowing for an earlier transfer. Earlier transfers may be preferable as selection at blastocyst stage can lead to loss of embryos during extended culture and can be a source of adverse health effects during and after pregnancy. As an example, a UK study using data from the HFEA register reported that while blastocyst transfer showed an increased odds ratio of 1.4 for live birth, it also showed an increased risk of the same magnitude for pre-term birth[31] . These disadvantages are not unexpected, since blastocyst culture acts to expose embryos to selection stress in a suboptimal in vitro environment, at the precise time in development when the embryonic genome and epigenome are being reset as part of the formation of a new individual[1,34,42]. Thus, there are strong clinical reasons for developing methods to select embryos for transfer and freezing at an earlier stage of development such as day 3[43]. In addition, embryo selection at even earlier stages (e.g. pronuclei on day 1) would have advantages in terms of rationing resources and space in clinical laboratories, and in specific clinical scenarios such as elective freeze-all of embryos in cases of ovarian hyperstimulation syndrome[44] .

Embryo assessment, both manual and ML, typically has an emphasis on specific chosen timepoints, however these may not necessarily be the optimal moments of embryo development for assessing viability. Here we applied eMLife to select the key developmental time-points that produced the best live birth predictions. We noticed that the peaks in model performance found at PN stage and 8-16 cell +21hrs are both just before certain well described developmental events; NEBD and blastocyst cavitation respectively. There are two possible explanations for this; at these pre-event time-points there is less natural variation in the appearance of viable embryos (compared to time-points where the embryo is undergoing processes such as PN growth or compaction) so it is easier to distinguish important developmental abnormalities, or these may be biologically important moments where any deviation from normal development can prevent the embryo from developing properly. For example, the ability of embryos to correctly prepare and execute the first mitosis is one of the defining moments of development and any anomalies around this time, such as problems with the NEBD process, can result in the developmental failure. The peak during the transition from the 4 to 8 cell stage coincides with embryonic genome activation in human embryos[45,46], raising the possibility that there are some morphological manifestations that can indicate the successful activation of the genome.

Additionally, we have provided evidence that predictions from early development may be used in conjunction with predictions from the blastocyst stage to give a better assessment of embryo quality than the blastocyst model alone. Further investigation into combining multiple model outputs and optimising decision thresholds would be an exciting avenue for future studies.

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

Our stage classification models also showed that MobileNetV2 pre-trained on ImageNet can give high accuracies on our dataset even with very small amounts of training data for simple classification tasks. Other studies have also used ImageNet for classifying human embryos[4,9,10,12,13,17-19,21,24], however our work is the first to investigate the performance that can be achieved with very little embryo data. Our data is quite different from the typical images included in ImageNet, which is mainly composed of photographs of everyday objects and animals rather than medical images. Therefore, it is useful to see that the fixed high level features obtained from ImageNet can be used successfully in this context. This finding may be relevant to IVF clinics wishing to use a small number of embryos to develop ML algorithms for purposes such as automating frame selection or detecting specific but rare abnormalities. Additionally, the ability of our stage classification models to obtain high accuracies even with very limited amounts of data could also be of interest for a wide range of biological and medical applications as it demonstrates that machine learning, and deep learning in particular, does not always require huge datasets.

Conclusion

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

Material and methods

1. Patients and time-lapse videos

Time-lapse videos of developing embryos were supplied by the IVF clinic in the Department of Reproductive Medicine at Old Saint Mary's hospital, Manchester University NHS Foundation Trust, Manchester, UK. Embryos were cultured in either Embryoscope ™ or Embryoscope+ ™ time-lapse incubator system (Vitrolife, Sweden) at 37°C, 6% CO_2 and 5% O_2. Throughout the study period there were no changes to the culture media or other culture conditions. Embryos were cultured in GTL overlaid with Ovoil (Vitrolife, Sweden) from post-injection (after ICSI, approximately 40 hours post-hCG trigger) to day 5 of development. Embryoscope ™ slides hold up to 12 embryos in individual wells holding 25µL of media, overlaid with 1.2ml oil. Embryoscope+ ™ slides hold up to 16 embryos, in 2 rows of 8 wells, each row of 8 wells is overlaid with 180uL of media. Embryos were not removed from the incubators during the observation periods, and media was not changed. Each time-lapse video exported had a framerate of 5-10 frames an hour. All images were irreversibly anonymised by clinic staff before being given to us.

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

2. Preparation of input image data

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

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

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

3. Model

The model we have used is the MobilenetV2 model[37] with weights pre-trained on the ImageNet database[47]. We used fixed convolutional features, only training the final layer of the model. All developmental stage and initial transfer outcome training attempts used a base learning rate (BLR) of 0.0001, a drop out of 0.5 and the cross entropy loss function. For the transfer outcome models we applied a class weight of 2 to the live birth class to account for this class having about half the amount of data as the no pregnancy class.

For the transfer outcome models we also experimented with using multi-class stage classification for extra transfer learning. The first step was to add an extra hidden layer to the MobileNetV2 model and train it to predict developmental stage with both the hidden layer and last layer trainable (the rest of the network was fixed as before). In the 2nd step of training we then took this model as a starting point for our outcome prediction models, for this step the only trainable layer was the last layer, so all the convolutional layers and the hidden layer were fixed. We repeated these two steps with various numbers of hidden units (100,320,640, or 960) in the hidden layer to find the optimal model structure. For stages earlier than the blastocyst we used the classes as in Fig. 4E for step 1. For the blastocyst stage we trained a separate model with the before NEBD class replaced with a blastocyst class and all images uncropped so that the resolution was the same at all stages.

4. Model training

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

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

5. Model performance evaluation

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

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

For the blastocyst model, we used the LIME software[41] so investigate how the model was choosing to classify blastocyst images. LIME (short for Local Interpretable Model-agnostic Explanations) uses perturbed images to understand how a model is making predictions. Taking in a model and a single image as an input, LIME produces output images that show the parts (known as super-pixels) of the original image that were used to classify it.

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

18 Fitz, V. W. *et al.* Should there be an "AI" in TEAM? Embryologists selection of high implantation potential embryos improves with the aid of an artificial intelligence algorithm. *J Assist Reprod Genet* **38**, 2663-2670 (2021). https://doi.org:10.1007/s10815-021-02318-7

19 VerMilyea, M. *et al.* Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Hum Reprod* **35**, 770-784 (2020). https://doi.org:10.1093/humrep/deaa013

20 Diakiw, S. M. *et al.* Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. *Human reproduction (Oxford)* **37**, 1746-1759 (2022). https://doi.org:10.1093/humrep/deac131

21 Payá, E., Bori, L., Colomer, A., Meseguer, M. & Naranjo, V. Automatic characterization of human embryos at day 4 post-insemination from time-lapse imaging using supervised contrastive learning and inductive transfer learning techniques. *Computer methods and programs in biomedicine* **221**, 106895-106895 (2022). https://doi.org:10.1016/j.cmpb.2022.106895

22 Miyagi, Y., Habara, T., Hirata, R. & Hayashi, N. Feasibility of predicting live birth by combining conventional embryo evaluation with artificial intelligence applied to a blastocyst image in patients classified by age. *Reprod Med Biol* **18**, 344-356 (2019). https://doi.org:10.1002/rmb2.12284

23 Ueno, S. *et al.* Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: a single-center large cohort retrospective study. *Fertil Steril* **116**, 1172-1180 (2021). https://doi.org:10.1016/j.fertnstert.2021.06.001

24 Sawada, Y. *et al.* Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. *Reprod Biomed Online* **43**, 843-852 (2021). https://doi.org:10.1016/j.rbmo.2021.05.002

25 Roberts, S. A., Hirst, W. M., Brison, D. R. & Vail, A. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. *Human reproduction (Oxford)* **25**, 2792-2802 (2010). https://doi.org:10.1093/humrep/deq213

26 Alfarawati, S. M. S. *et al.* The relationship between blastocyst morphology, chromosomal abnormality, and embryo gender. *Fertility and sterility* **95**, 520-524 (2011). https://doi.org:10.1016/j.fertnstert.2010.04.003

27 Ziebe, S. *et al.* FISH analysis for chromosomes 13, 16, 18, 21, 22, X and Y in all blastomeres of IVF pre-embryos from 144 randomly selected donated human oocytes and impact on pre-embryo morphology. *Human reproduction (Oxford)* **18**, 2575-2581 (2003). https://doi.org:10.1093/humrep/deg489

28 Victor, A. R. *et al.* One hundred mosaic embryos transferred prospectively in a single clinic: exploring when and why they result in healthy pregnancies. *Fertility and sterility* **111**, 280-293 (2019). https://doi.org:10.1016/j.fertnstert.2018.10.019

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

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

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

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

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

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

35 Deshpande, P. & Gupta, A. Causes and prevalence of factors causing infertility in a public health facility. *Journal of human reproductive sciences* **12**, 287-293 (2019). https://doi.org:10.4103/jhrs.JHRS_140_18

36 Castillo, C. M. *et al.* The impact of IVF on birthweight from 1991 to 2015:a cross-sectional study. *Human reproduction* **34** (920-931) (2019).

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

38 Omer, S. & Lior, R. Ensemble learning: A survey . *WIREs Data Mining Knowl Discov.* **8**, e1249 (2018).

39 Gardner, D. K., Lane, M., Stevens, J., Schlenker, T. & Schoolcraft, W. B. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertility and sterility* **73**, 1155-1158 (2000). https://doi.org:10.1016/S0015-0282(00)00518-5

40 Balaban, B. *et al.* Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Reproductive biomedicine online* **22**, 632-646 (2011). https://doi.org:10.1016/j.rbmo.2011.02.001

41 Ribeiro, M.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the NAACL-HLT 2016–2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session,* San Diego, CA, USA, 12–17 June 2016.

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

43 Marconi, N., Allen, C. P., Bhattacharya, S. & Maheshwari, A. Obstetric and perinatal outcomes of singleton pregnancies after blastocyst-stage embryo transfer compared with those after cleavage-stage embryo transfer: a systematic review and cumulative meta-analysis. *Human reproduction update* **28**, 255-281 (2022). https://doi.org:10.1093/humupd/dmab042

44 Papanikolaou, E. G. *et al.* Incidence and prediction of ovarian hyperstimulation syndrome in women undergoing gonadotropin-releasing hormone antagonist in vitro fertilization cycles. *Fertility and sterility* **85**, 112-120 (2006). https://doi.org:10.1016/j.fertnstert.2005.07.1292

45 Braude, P., Bolton, V. & Moore, S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature (London)* **332**, 459-461 (1988). https://doi.org:10.1038/332459a0

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

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

Supplementary materials

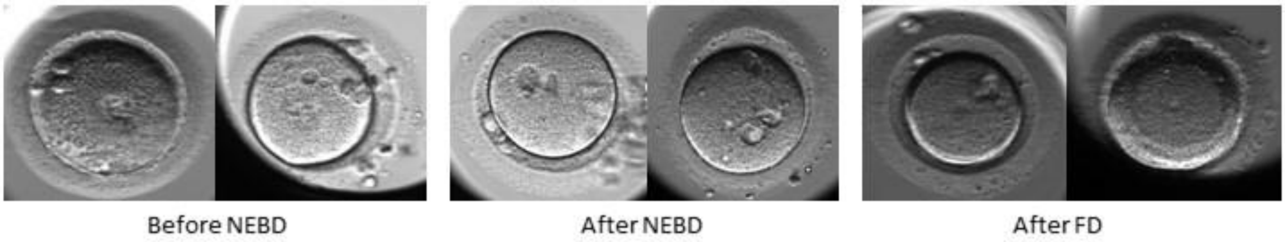
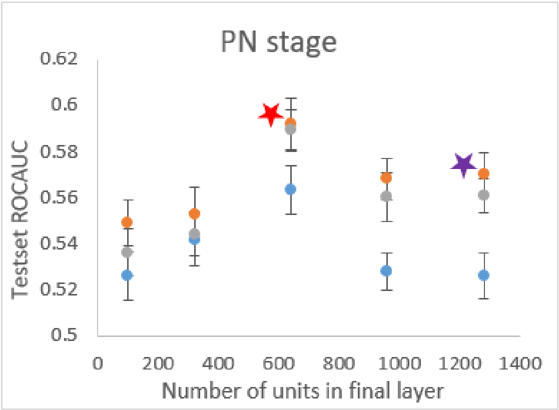
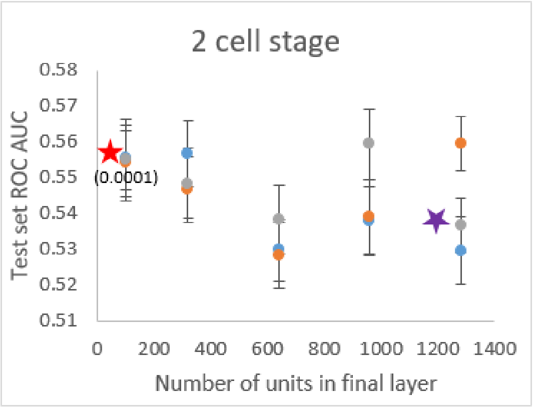
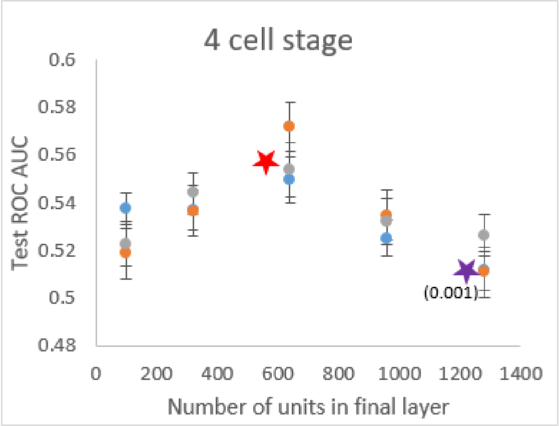
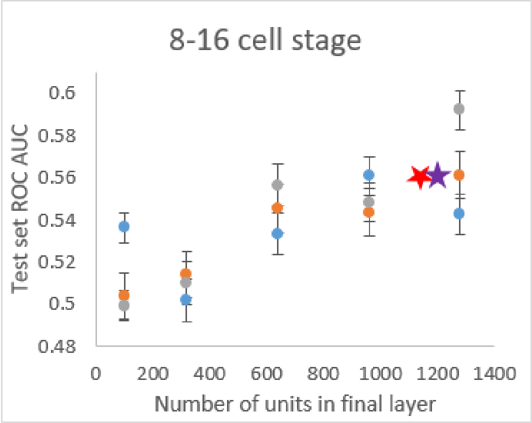


Figure S1A: Examples of images incorrectly classified by stage selection models. Left: images before NEBD where the PN are hard to see. Middle: images without PN with vacuoles that resemble PN. Right: 2 cell embryos that look like 1 cell embryos.









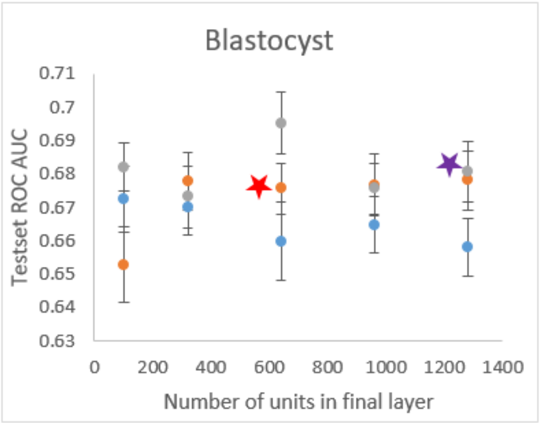


Figure S2A: Hyper-parameter tuning for each stage. The models with 1280 hidden units use the standard MobileNetv2 model with no extra hidden layers. All other amounts of hidden units use an extra hidden layer that had been pre-trained on the stage classification dataset. The red star marks the best model set up overall for the validation set and the purple star marks the best model set up overall for the validation set when the standard model is used

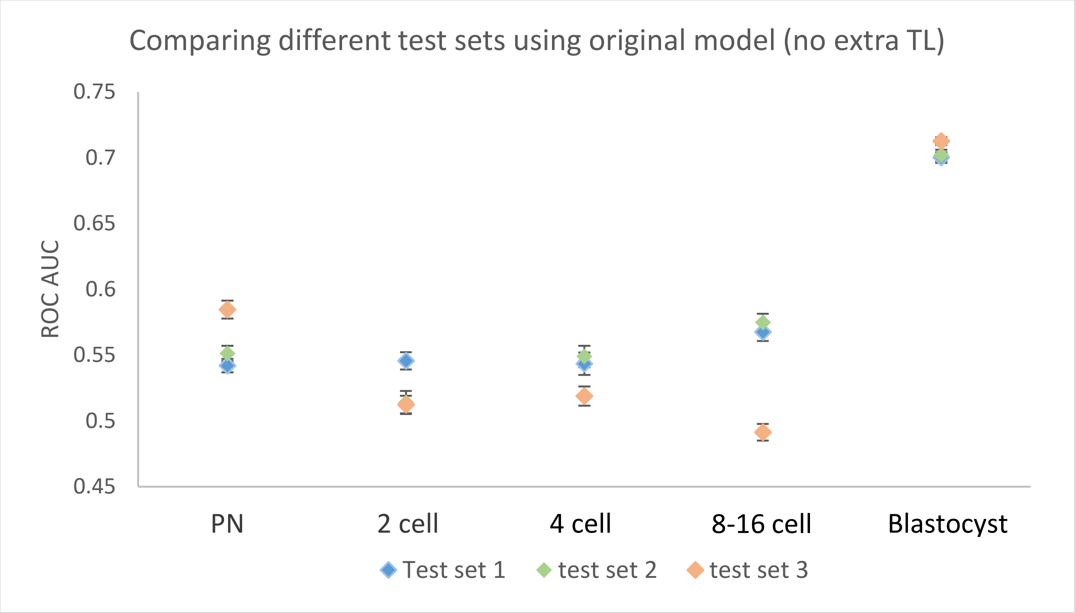


Figure S2B: Comparison of the ROC AUC obtained by 3 different hold out test sets at each stage. A BLR of 0.0001 and the original transfer learning model with no extra transfer learning was used.

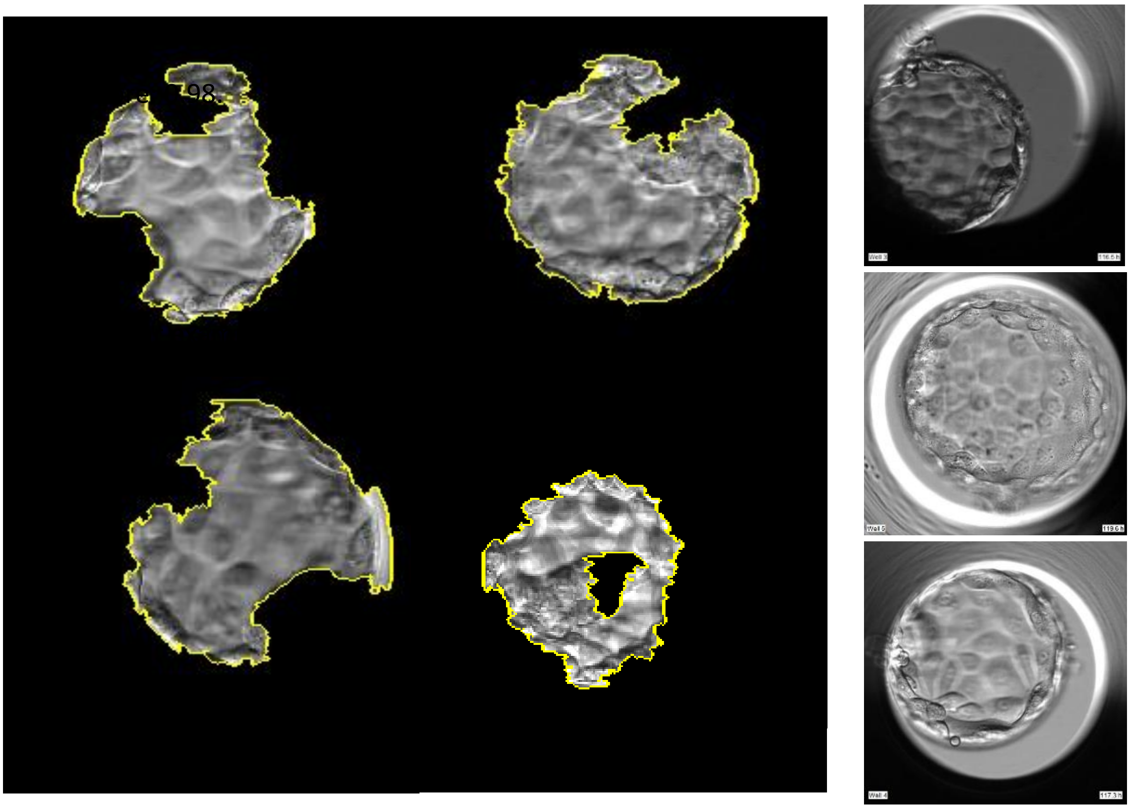


Figure S3A: Investigating the Blastocyst model. Left: LIME explanations of the blastocyst model. The pictures show the areas of the image that suggested to the model that the image was in the positive (live birth) class. Right: Examples of blastocyst images that were given a very high blastocyst model score (>0.98) despite the ICM not being in focus, or in some cases not visible at all.

Chapter 3:

Combined deep learning and human annotation approach allows cleavage stage assessment of human embryos

Camilla Mapstone, Helen Hunter, Isobel Green, Riannah Hayes, Blessing Oderinde, Daniel Brison, Julia Handl, Berenika Plusa

Rationale of paper

Following the successful development of CNN models predicting live birth in chapter 2, we next wanted to investigate whether the predictions from these models could be combined to further improve live birth prediction at pre-blastocyst stage. We also wanted to investigate whether manually annotated features might be able to contribute to live birth prediction. This paper is a draft manuscript that we intend to submit later this year. All supplementary material is included at the end of the manuscript.

Aims

-Identify morphological features useful for embryo viability assessment

-Use machine learning algorithms to predict live birth using only CNN predictions and morphological features from cleavage stage and earlier.

Author contributions

Dr Berenika Plusa, Dr Julia Handl and myself contributed to the concept of the study. Helen Hunter collected the time-lapse videos from the IVF clinic. Dr Berenika Plusa and myself designed the methodology for assessing the morphological features included in the screening stage, with measurements and qualitative assessments carried out by myself and Masters students (Isobel green, Riannah Hayes, and Blessing Oderinde) supervised by me. I carried out subsequent data analysis and produced graphs. I performed all dimensionality reduction and unsupervised and supervised modelling and analysed the results. The manuscript was drafted by myself and then critically reviewed by Dr Berenika Plusa and Dr Julia Handl.

Combined deep learning and human annotation approach allows early assessment of human embryos.

Camilla Mapstone[1,2], Helen Hunter[3], Isobel Green[1], Riannah Hayes[1], Blessing Oderinde, Daniel Brison[3,4], Julia Handl[2], Berenika Plusa[1]

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

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

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

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

Abstract

Embryo assessment is one of the main challenges faced by IVF and is currently prone to high levels of subjectivity. An automated algorithm combining deep learning with manually assessed morphological features may be able to improve the accuracy and objectivity of the embryo selection process. We have investigated various potential morphological markers of embryo viability and identified a few that are predictive of live birth. These markers were then combined with deep learning predictions to give an improved overall prediction of live birth at pre-blastocyst stage. Additionally, we analysed the dataset using dimensionality reduction techniques and unsupervised learning, giving a deeper insight into the relationship between embryo morphology and transfer outcome.

Introduction

Infertility has become a highly prevalent medical issue, affecting 8-12% of couples worldwide[1]. In vitro fertilisation (IVF) is a common treatment for infertility and has been rapidly rising in popularity since the first IVF birth in 1978[2]. In recent years single embryo transfer (SET) has been favoured to mitigate the potential risks associated with multiple pregnancies[3-5], therefore it is very important that the viability of the embryos can be assessed accurately to increase the chance of selecting an embryo that will led to a healthy live birth. However, the task of selecting an embryo to transfer is challenging as it is still not fully understood what constitutes healthy embryo development[6-8].

The introduction of time-lapse monitoring has resulted in an increase in the amount of information available on embryo development as the embryos are recorded continuously[9]. As a result, various morphological and kinetic features have been studied throughout development, such as PN dimensions[10-14], cytoplasm appearance[15-17], and cell divisions[18-22], and these have been linked to outcomes such as blastocyst formation, clinical pregnancy, implantation, or live birth[16,23-29] with various degrees of success. Moreover, there are still some morphological features for which the link to embryo viability is unclear due to conflicting reports, such as zona pellucida (ZP) thickness variation[30-32] and the presence of vacuoles at zygote stage[10,16].

It is thought that machine learning (ML) may be particularly well suited to the task of combining various features together to produce algorithms that can support assessment of embryo quality[33], and various ML approaches have been attempted to address embryo selection[33,34]. Previous efforts with ML either use manually extracted features, employ deep learning (DL) algorithms on image data or combine both approaches. It is possible that a combined DL/manual annotation approach would offer the best solution as it could capture a broad range of features that might otherwise be missed.

Many existing studies are based on assessment at the blastocyst stage as culturing until blastocyst has become common practice due to higher live birth rate per transfer[35]. It is easier to assess embryo quality at the blastocyst stage as the less viable embryos may fail to form the expected structure. However, blastocyst culture does not improve overall live birth success rate per cycle if all embryos are eventually transferred[35] and it is associated with some adverse health effects; mono-zygotic twinning, altered birthweight, pre-term birth, and shortened telomeres[35-39]. Therefore, an improvement in our ability to assess the quality of pre-blastocyst embryos could help to move towards earlier, and therefore safer, embryo transfers.

Here, we aimed to combine pre-blastocyst stage predictions from DL models trained in an earlier study (Mapstone et al., manuscript under submission to npj Digital Medicine)[40] with manual annotations to get an improved assessment of early stage embryos, allowing for earlier transfer and more diversification in treatment regimes . First, we used a dataset of time-lapse videos of embryos with various clinical outcomes to investigate a number of morphological features and identify those that may be predictive of live birth. Next, we used both dimensionality reduction and K-means clustering to gain a deeper understanding of our dataset. Finally, we experimented with a range of supervised learning models to predict live birth, and found that combing manual features with predictions from DL models gave improved performance.

Results

In this study we aimed to combine predictions from previously developed DL models[40] with manually extracted features to get an improved early stage prediction of live birth following embryo transfer. The data used in this study were time-lapse videos of ICSI embryos from 2016-2019 provided by Saint Marys hospital, Manchester. Embryos with three different outcomes were included; transferred and resulted in live birth (LB), transferred and did not result in pregnancy (NP), and discarded due to poor quality. For our final model we used only the NP and LB groups as we wanted to be able to differentiate between embryos that are chosen to be transferred to stretch the boundaries of human decision making.

Investigation into potential quality markers

We started with a screening stage in which we used all three outcome groups to identify potential viability markers. This involved investigating the link between many different morphological features and embryo outcome in a subset of the data; all embryos from 2016 (157 in total). A subset was used due to time constraints and embryos from 2016 were chosen as our dataset included the complete set of ICSI embryos for this year so could ensure all transfer outcome groups were from the exact same time period. Features were chosen either due to having previously been linked to embryo quality in the literature, or because a preliminary examination of the time-lapse videos suggested to us that the feature might be useful. We concentrated mainly on the pre-blastocyst stage to support early viability assessment. The features investigated were divided into two groups: quantitatively measured features and qualitative features.

The results of the quantitative features are shown in Fig 1a along with a timeline of embryo development illustrating the developmental stage at which the measurements were taken. The quantitative features measured were:

·PN area: $\pi(\frac{PNL_1+PNL_2}{4})^2$

·PN area difference: $\pi(\frac{PN1L_1+PN1L_2}{4})^2 - \pi(\frac{PN2L_1+PN2L_2}{4})^2$

·ZP thickness: $\frac{ZP_1+ZP_2+ZP_3+ZP_4}{4}$

·ZP thickness variation: $ZP_{max} - ZP_{min}$

·Duration of 3 cell stage: Time of first frame with 4 cells – Time of first frame with 3 cells

·Average compaction angle: $(\sum_{k=0}^{n} \mathrm{CA}^k)/n$ where CA is compaction angle and n is number of angles measured.

·Perimeter difference: Perimeter of morula - Circumference of enclosing circle

·ZP thickness difference: $\frac{ZPZyg_1+ZPZyg_2+ZPZyg_3+ZPZyg_4}{4} - \frac{ZPblast_1+ZPblast_2+ZPblast_3+ZPblast_4}{4}$

At the zygote stage we investigated PN size as differently sized PN has previously been linked to live birth[11] and ZP thickness to address currently conflicting reports[30-32]. Each PN was measured twice, with measurements taken parallel with and perpendicular to the edge touching the other PN (see Fig. S1a). The area was then calculated by approximating the PN shape as a circle and taking the average of these two measurements as the diameter. The average PN area and the difference in area between the two PN were then calculated for each embryo and these values were compared between LB, NP and discarded groups (Fig. 1a). The results showed no statistical difference in either of these measurements. Next, we assessed the thickness of the ZP by taking measurements in four fixed locations, as illustrated in Fig. S1a, with measurements skipped if the required part of the ZP was obscured by the side of the dish. The average thickness and thickness variation, defined as: (max measurement – min measurement), were calculated from all the measurements that could be taken, embryos with less than two measurements were not included. The thickness was found to be significantly higher in the discarded group than the LB group but did not significantly differ between NP and LB or NP and discarded. The thickness variation values were not found to differ significantly between any different outcomes.

Although the features analysed at zygote stage did not allow for distinction between the LB and NP groups, we were interested to see whether a combination of features might be more useful. To reduce this 4-dimensional space down to an easier to visualise 2-dimensional space we used both 2-component PCA and 2-component t-SNE with varying levels of perplexity, results are shown in Fig. S1b-c. For both PCA and t-SNE there does not seem to be much difference between the NP and LB groups. For PCA, the discarded group also does not appear to differ, however the t-SNE plots do appear to show a region of the feature space that has less discarded embryos, and this is conserved throughout varying perplexity values. Neither PCA or t-SNE appear to produce clearly distinct clusters or groupings of the data (except low perplexity t-SNE, which is likely to be an artefact of the process). Overall, the zygote stage features do still not appear to show any clear separation between outcome groups even when combined.

Next, we recorded the 3-cell stage duration (reflecting the asynchrony between the 2nd and 3rd mitosis) as longer durations have been negatively linked to blastocyst formation[19]. The durations were found by subtracting the number on the timestamp in the first frame with three cells from the number on the timestamp in the first frame with four cells. The results (Fig. 1a) show that there was little difference between the LB and NP group, however the discarded group was in the 3 cell state for a statistically significant longer time period than both the LB and NP groups. However, it should be noted that it was often difficult to observe exactly when cell divisions occurred in the discarded group as the cleavage stage frequently exhibited a lot of fragmentation and false divisions. Our results suggest that although the discarded embryos can be identified using this feature, division asynchrony is very similar between the embryos chosen for transfer.

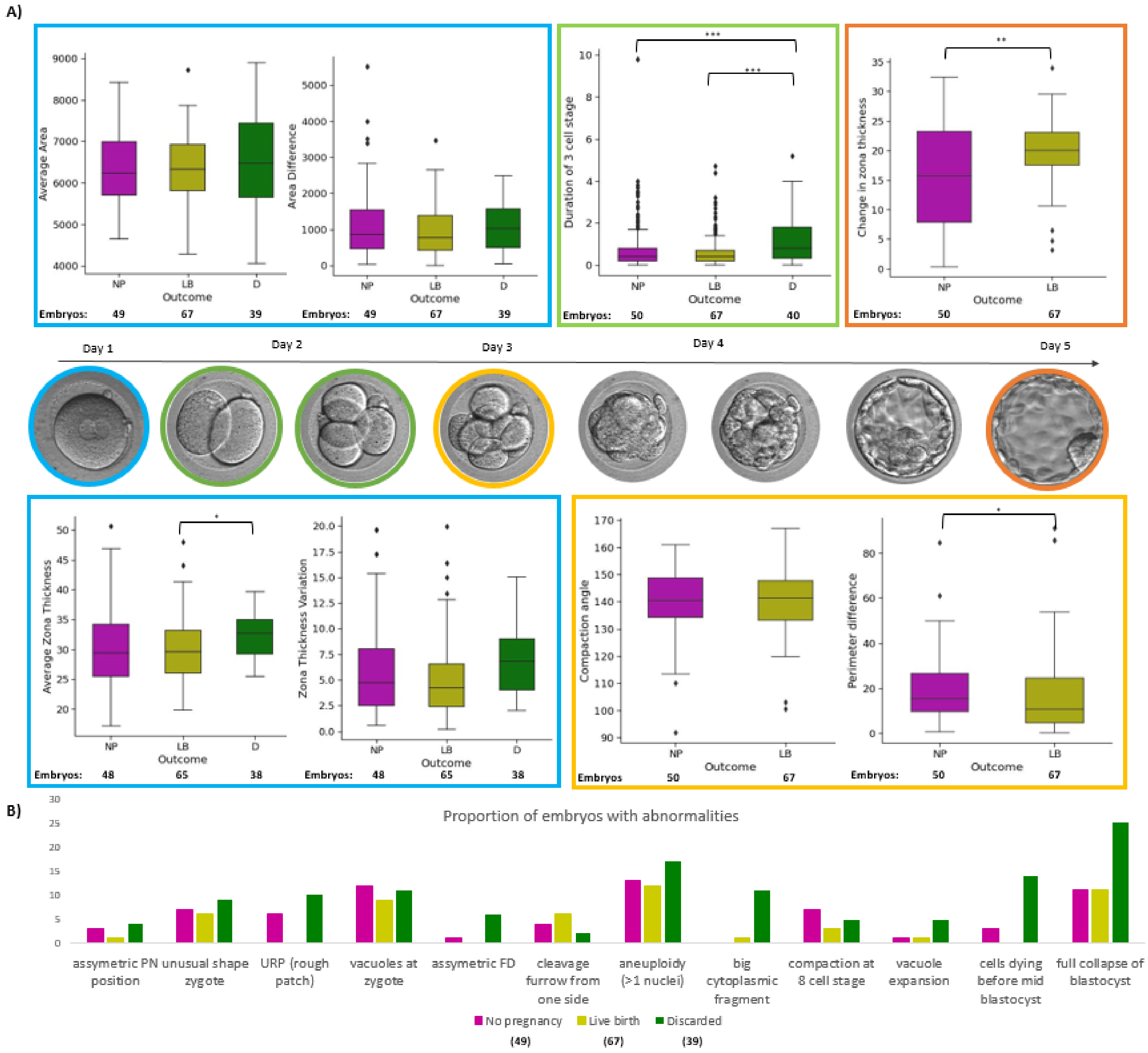
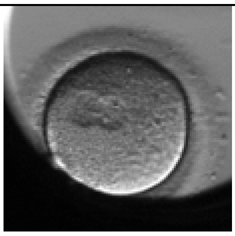
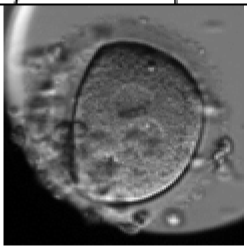
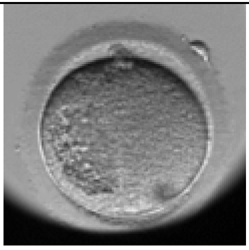
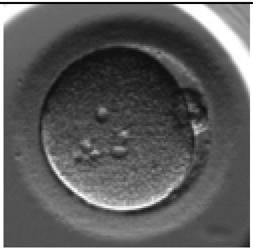
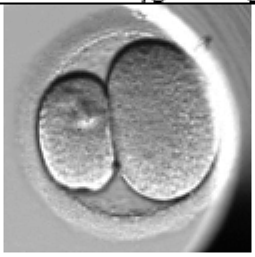
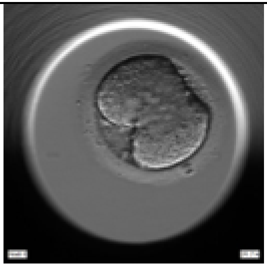


Figure 1: Correlations between various morphological features and transfer outcome. A) Distribution of values for various quantitative features for the LB, NP and discarded outcome groups. The developmental stage that each feature was measured at is illustrated by the images in the embryo development timeline. B) Qualitative features/'markers' in the LB, NP, and discarded transfer outcome groups. The number of embryos with the marker present is shown for each marker and each transfer outcome group.

Qualitative feature	Description	Source
 Asymmetric PN position	PN still not in a central position just before NEBD	Observation from preliminary data examination
 Unusual zygote shape	Zygote shape that deviates from the typical circular appearance	Observation from preliminary data examination
 Rough patch (RP)	Area of unusual texture at zygote stage	Observation from preliminary data examination
 Vacuoles at zygote stage	One or vacuoles present at the zygote stage	*Occurrence and developmental consequences of vacuoles throughout preimplantation development*[16] (Ebner et. al 2005)
 Asymmetric FD	Very uneven cell sizes immediately after FD	*The use of morphokinetics as a predictor of implantation: a multicentric study to define and validate an algorithm for embryo selection*[23] (Basile et al 2015)
 Cleavage furrow from one side	Asymmetric cleavage furrow during FD	Observation from preliminary data examination

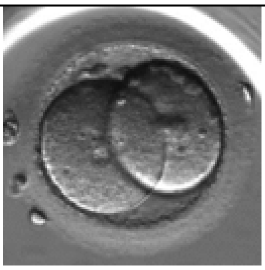
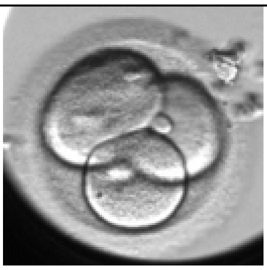
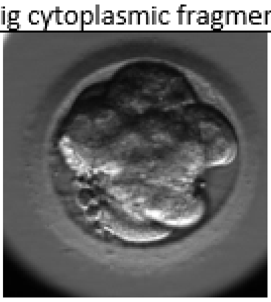
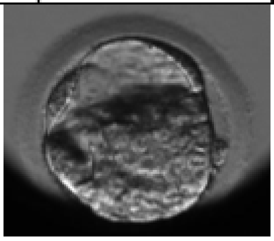
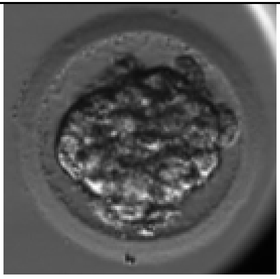
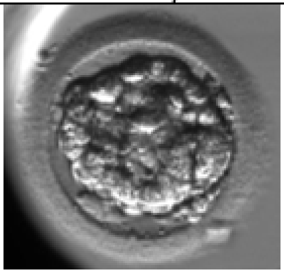
 Multi-nucleation	More than one nuclei in one or more cell at 2 cell stage	*Multinucleation of a sibling blastomere on day 2 suggests unsuitability for embryo transfer in IVF-preimplantation genetic screening cycles*[48] (Ambroggio et al 2011) and *Multinucleation in cleavage stage embryos*[49] (Van Royen et al 2003)
 Big cytoplasmic fragment	Unusually large cytoplasmic fragment – only those of a comparable size to a cell included	*The use of morphokinetics as a predictor of implantation: a multicentric study to define and validate an algorithm for embryo selection*[23] (Basile et al 2015)
 Compaction at 8 cell stage	Very early compaction – starting before 8 cell stage	*Analysis of compaction initiation in human embryos by using time-lapse cinematography*[24] (Iwata et al 2014)
 Vacuole expansion	A vacuole expands at morula stage, often in place of cavitation.	*Occurrence and developmental consequences of vacuoles throughout preimplantation development*[16] (Ebner et. al 2005)
 Cells dying before mid-blastocyst	Cells seen to be dying before the embryo reaches mid-blastocyst stage	Observation from preliminary data examination
 Full collapse of blastocyst	The expanding blastocyst undergoes one or more collapse where the cavity completely disappears temporarily.	*Human blastocyst spontaneous collapse is associated with worse morphological quality and higher degeneration and aneuploidy rates: a comprehensive analysis standardized through artificial intelligence*[50] (Cimadomo et al, 2022)

Table 1: Catalogue of qualitative features investigated

At the morula stage, we attempted to measure the roundness of the embryo at the time-point where it appeared most compacted. This moment of development was chosen because our previous work had suggested it may be useful for predicting live birth[40]. Only the NP and LB groups were used for this analysis as many of the discarded embryos did not reach this stage. We first used an 'angle method' previously used by Maître et al. (2015)[41] to assess degree of compaction in mouse embryos. The angle method found the average angle between cells on the outside of the embryo, with larger angles suggesting a more compacted embryo (See Fig. S1a). We did not find a significant difference in average angle between NP and LB, however we found this method difficult to perform on many embryos as it was often hard to judge where the angles should be measured, which could have influenced the accuracy of the results. Therefore, we next tried an approach we called the 'perimeter method', which calculated the magnitude difference between the length of the perimeter of the embryo and the circumference of the smallest enclosing circle, a smaller number indicates a rounder/ more compact embryo (see Fig S1a). We found this method to be less subjective than measuring angles. The perimeter method found a significant difference; the NP group had a larger difference between perimeter and circle circumference than the LB group suggesting that NP embryos were slightly less round/compacted on average. Although only the perimeter method produced significant results, both methods indicated that on average there were rounder/ more compacted embryos in the LB group (the average angle was slightly larger in the LB group), so the perimeter method may be a more reliable method for measuring compaction.

In addition, as ZP thinning is an established component of embryo grading[42], ZP measurements were taken at blastocyst stage using the same method as used previously for the zygote stage and the change in thickness between zygote and blastocyst stage was calculated for each embryo. Again, this was only carried out for the NP and LB groups as many of the discarded embryos did not form blastocysts. The results (Fig. 1a) show that a significantly bigger change was seen in the LB group, as might be expected due to this group being more likely to include properly developed blastocysts that would push against the ZP during expansion leading to ZP thinning[43].

Next, we investigated a range of qualitative features from throughout development that we thought may be indicative of live birth. Descriptions and illustrations of each feature are shown in table 1 along with the reason for inclusion. These features were all specific morphological phenomena, that we refer to here as 'markers', and were assessed in a binary fashion; each marker was recorded as either present or absent for each embryo. Fig. 1B shows the proportion of embryos with each marker in each group. We found that the difference was most noticeable between the LB and discarded group, with the discarded group showing higher proportions of every marker except cleavage furrow from one side. The difference between the LB group and the NP was less strong, however the NP group still showed a greater proportion of embryos with the marker present for 9 out of 12 markers (all except cleavage furrow from one side, vacuole expansion and full collapse of blastocyst).

Our screening stage analysing the dataset of embryos from 2016 has highlighted a few potential differences in the morphology of embryos with different outcomes. As expected, this was more strongly marked when comparing the LB to the discarded group, however there were still noticeable differences between the LB and NP groups, Suggesting the potential to improve upon human decision making.

Full dataset analysis

Following the screening stage using the 2016 data sub-set, we selected some features to look into further with the full dataset. We chose features for which there appeared to be a difference amongst the NP and LB groups: PN asymmetric location, rough patch (RP), zygote vacuoles, uneven FD, and large cytoplasmic fragment. As we were interested in early stage prediction we only used features up to the cleavage stage.

The results for all three outcome groups for each feature are displayed in Fig. 2a. We can see that the discarded group has the highest frequency of all markers and the LB group has the lowest. The difference between the discarded group and the LB group is significant for all qualitative features apart from asymmetric PN position, while the difference between the discarded group and NP group is just significant for asymmetric FD and big cytoplasmic fragment. The differences between the LB and NP group are all of low-significance (<0.05, but>0.01) or not significant, however the markers were always more frequent in the NP group. The fact that much higher frequencies were seen for all markers in the discarded group, which all failed to form blastocysts, than the LB group adds evidence that these markers are correlated with lower developmental potential.

Next, we combined the NP and LB qualitative features data with predictions from DL models (developed in an earlier study) trained to predict live birth at the PN, 2 cell and 4 cell +14 hours stage. These models were all CNNs that had been trained using a transfer learning approach. The combination of the CNN predictions and qualitative features resulted in a dataset with 8 features in total. The discarded group was not included as outcomes are not known for certain for this group since these embryos were not transferred, and this dataset was created for developing live birth prediction models. Qualitative features were all binary (1 if the marker was present, 0 otherwise) and model scores were originally a continuous number from 0 to 1. As having a combination of binary and non-binary features could possibly make further analysis difficult we also created a fully binary dataset by assigning a value of 1 when the model score was <0.1 (so in theory corresponding to poor embryo quality) and a 0 otherwise.

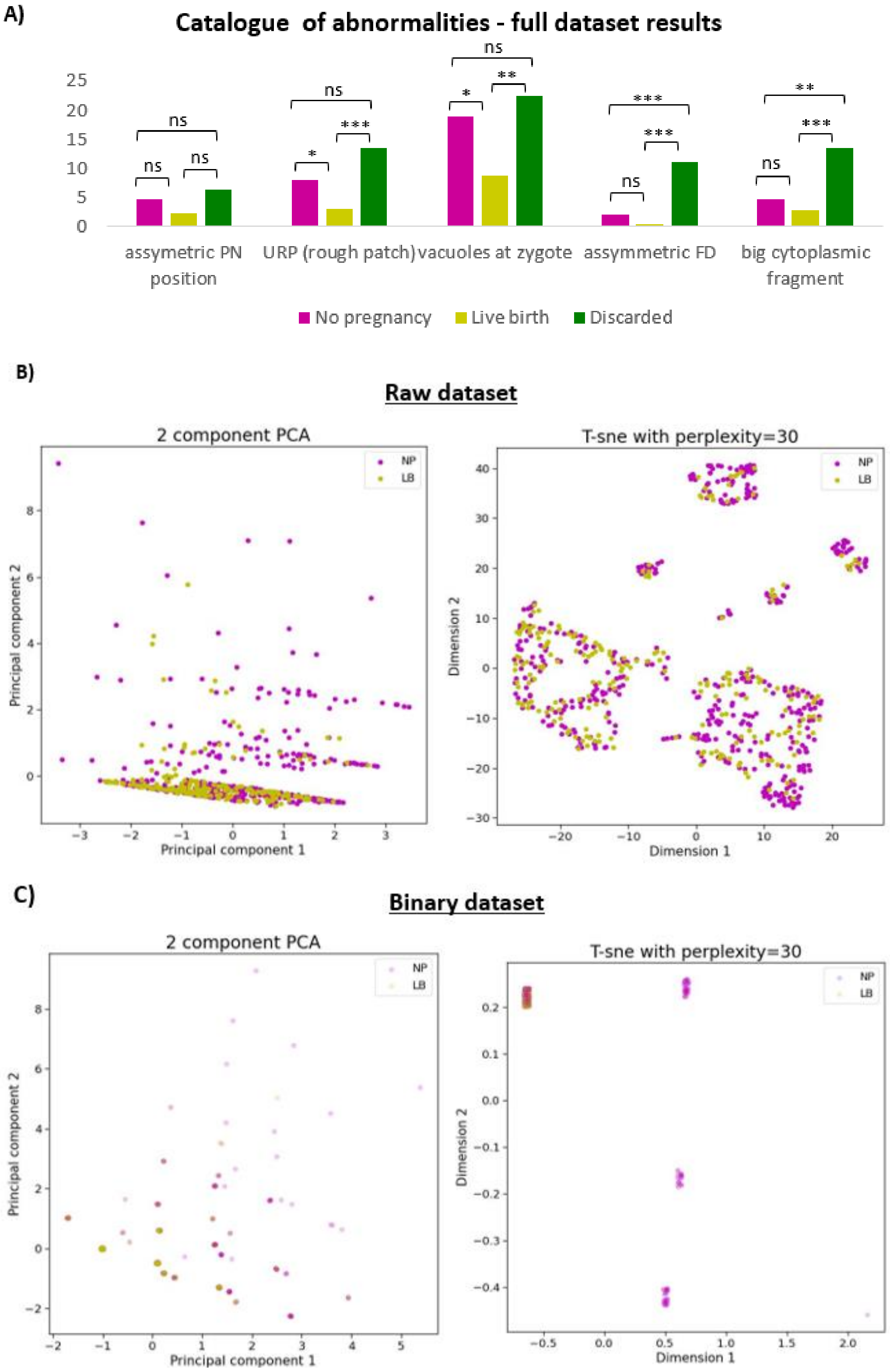


Figure 2: Final dataset features and visualisation. A) Qualitative features/'markers' in the LB, NP, and discarded transfer outcome groups that have been chosen for inclusion in the final dataset. The number of embryos with the marker present is shown for each marker and each transfer outcome group. B) visualisations of the raw dataset using 2-component PCA and t-SNE. C) visualisations of the raw dataset using 2-component PCA and t-SNE.

We then carried out 2-component PCA and 2-component t-SNE with default parameters for both the binary and non-binary datasets for the NP and LB groups to visualise the data, the results are shown in Fig. 2b. For the binary dataset there are many embryos with the exact same feature values so the data markers were made translucent and randomly offset by up to 0.02 using jittering on each axes so that overlapping data-points could be seen. Explained variance ratio for all PCA components are shown in tables 2a and 2b for non-binary and binary datasets respectively and the component loading is shown in tables 2c-d. It can be seen that the model scores account for more variance than the qualitative features in both datasets.

The PCA visualisation show the same general pattern for the binary and non-binary dataset, with differences in the NP and LB groups clearly apparent, despite a substantial amount of overlap. A large proportion of the data-points formed one large cluster with a mix of outcomes, while the remaining data did not appear to belong to any particular cluster. The t-SNE results appear to vary more between the binary and non-binary datasets. The non-binary dataset produced 2 large clusters that appeared to have a mix of NP and LB embryos and a few smaller clusters. The binary dataset produced one cluster of mixed outcomes and 4 other clusters of only NP embryos. We also checked how varying perplexity would affect our results (Fig. S2a-b), and the trends seen appear to be fairly robust but break down for lower perplexity values. This suggests that lower perplexity values may not be appropriate for this dataset as repulsion between data-points is too strong compared to the attraction between data-points so the natural clusters are broken up.

Component	Explained variance ratio
1	0.192
2	0.169
3	0.148
4	0.124
5	0.099
6	0.09
7	0.087
8	0.086

Table2a: PCA components for non-binary data

Component	Explained variance ratio
1	0.179
2	0.165
3	0.152
4	0.124
5	0.109
6	0.099
7	0.089
8	0.083

Table2b: PCA components for binary data

Feature	Component 1	Component 2
PN model	-0.53059	-0.04725
2 cell model	-0.54475	0.109026
4 cell model	-0.53133	0.134212
Asymmetric FD	-0.08423	0.526366
Asymmetric PN position	-0.02189	0.582906
Large cytoplasmic fragment	-0.16922	0.120141
RP	0.191686	0.427458
Vacuoles	0.25779	0.392267

Table 2c: Non-binary PCA component loadings

Feature	Component 1	Component 2
PN model	0.529957	-0.22679
2 cell model	0.408833	-0.27054
4 cell model	0.46248	-0.30339
Asymmetric FD	0.153871	0.52311
Asymmetric PN position	0.230162	0.54425
Large cytoplasmic fragment	-0.12688	0.187691
RP	0.269969	0.359058
Vacuoles	0.415306	0.222085

Table 2d: Binary PCA component loadings

As performing both PCA and t-SNE suggested the presence of one or more clusters in our dataset, we then investigated further by applying K-means with varying numbers of clusters (*k*). We first applied K-means to the original binary dataset to see what clusters would appear without any pre-processing. We visualised the results for each iteration by plotting 'success rate' of cluster against *k*, with the size of each point proportional to the size of the cluster it represented (Fig. 3). Success rate was defined as the fraction of embryos in the cluster that belonged to the LB group. The 'by chance' success rate is also shown by a dotted line – this is the success rate of the dataset as a whole, so is the expected success rate if the clusters had no correlation with outcome. The results show a distinct pattern; regardless of the number of clusters, there is always one cluster that is larger and more successful than all other clusters, whilst most of the other clusters are below the by-chance line. To check how much this pattern deviated from what might be expected by chance we also ran a control experiment where the outcome labels were randomly shuffled (Fig. S3a). This produced the expected pattern of larger clusters tending to be closer to the by-chance line and smaller clusters (more prone to statistical noise) randomly scattered both below and above the line. This confirms that the pattern we saw in the original experiment is different from what we would expect if outcome was unrelated to the features.

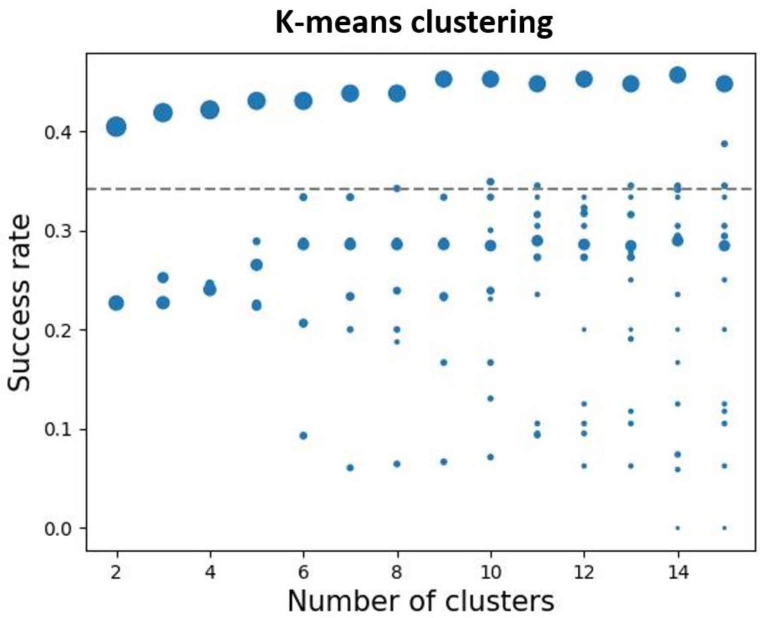


Figure 3: Clusters produced by K-means algorithm. A visualisation of the clusters produced when running K-means on the binary dataset for *k* values of 2 to 15, each marker represents a cluster. The success rate, defined as the fraction of embryos in the cluster that resulted in a live birth after transfer, is shown on the y axis. The dotted line shows the success rate for the dataset as a whole. The marker size is scaled proportionally to number of embryos in the cluster that it represents.

We then repeated clustering on the datasets produced by PCA and t-SNE to quantitatively investigate the clusters we had observed from the previous visualisations and confirm whether the same clustering pattern is still captured after dimensionality reduction. We used a varying numbers of components for PCA and varying values of perplexity for t-SNE (Fig. 3b-c) and found the results were similar to those seen with the unprocessed data. The trend became very exaggerated for t-SNE with a perplexity of 30 or above, with just one smaller cluster seen with 'success rate' of zero, presumably representing the clusters seen in the t-SNE visualisations that appeared to only have NP embryos.

Live birth prediction modelling

Finally, we were interested to see how well we could predict live birth using all of our features as inputs to supervised learning models. We kept all the qualitative features in the dataset even if they had not been shown to be significant as all differed between the NP and LB groups so could potentially still be useful when combined with other features by a model. We used both our binary and non-binary datasets on 4 different supervised ML models: linear regression, logistic regression, decision tree and support vector classification. We attempted this both with the raw feature values and after pre-processing the input data with PCA, trying component numbers ranging from 2 to 4. Models were trained 50 times with different train/test set random splits and the performance was evaluated by finding the average ROC AUC over all test sets. In Fig. 4 we show the performance of each model for the dataset we found to give the best performance; 3 component PCA on non-binary data. The ROC AUC (on this dataset) of the best performing pre-blastocyst stage DL model, the PN model, is shown with a dotted line as a baseline reference. The results show that linear regression appears to be giving the best performance, with a ROC AUC of 0.639, and decision tree the worst, although there is quite a bit of variance in the performance of repeat training attempts with different train/test splits.

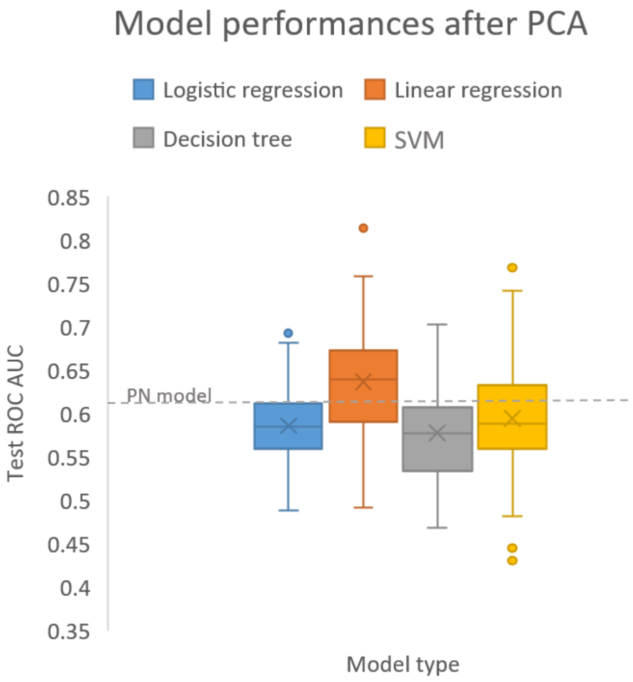


Figure 4: Performance of supervised ML algorithms trained to predict live birth. The test set ROC AUC values obtained by each model from 50 training attempts with random train/test splits. The ROC AUC of the best performing CNN model contributing predictions to the dataset, the PN model, is also shown as a baseline performance.

Full results for all input datasets and models are shown in table S1. We can see that PCA does generally appear to be making a small improvement to test set performance across all models. The difference between training set and test set performance is smaller when using PCA, and the training performance is slightly lower than what was achieved using raw features. This suggests that there is information loss due to PCA, however this has reduced overfitting and led to an overall improvement. We can also see from table S1 that better performance is generally achieved when using the non-binary rather than binary dataset, suggesting that information lost when converting the ML model output data to binary format does has some predictive value.

In summary, we have found that by combining our qualitative features with previously developed DL models we can get a better prediction of live birth than what was achieved by any of the DL models alone. We have also found that the best performing ML model for this task is linear regression, and pre-processing with PCA appears to slightly improve performance.

Discussion

There are many possible morphological features that could indicate the viability of an embryo, identifying potential markers of embryo quality and combining this information to make reliable predictions of live birth could improve IVF success rates. Here we have investigated several morphological features and identified those that could be useful for embryo viability assessment. We then used dimensionality reduction techniques and unsupervised learning to better understand our dataset. Finally, we experimented with a few different supervised learning techniques to predict live birth using only information up to the cleavage stage.

Our dataset included both transferred and discarded embryos. Many of the discarded embryos are clearly low quality so their inclusion in model development and testing would unfairly skew the results by including embryos that would never have been chosen for transfer in the 'unsuccessful' class. We also wanted to only use embryos with known outcomes for the model development, as the discarded group were not transferred and we do not know for certain that they all would have failed. However, we decided that it would be useful to include the discarded group in our analysis of potential quality markers as there is likely to be a stronger signal between the discarded group and the LB group.

Both measured and qualitative features had associated limitations that need to be taken into consideration. The ZP thickness and thickness variation values were based on sampled measurements and the area calculations for the PN assumed they were a circular shape. More accurate measurements via automated image processing techniques involving segmenting these areas is a potential avenue for further work. The qualitative features had the limitation that there is some level of subjectivity in deciding whether an embryo displayed each marker. If these qualitative features were to be used in practice it might be beneficial to train CNN models to detect each marker individually to get more consistent assessments.

The locations of PN and ZP measurements were pre-specified to ensure the methodology for taking measurements was consistent between groups. This did result in some embryos not being included as measurements could not be taken in the required positions, however we decided this was necessary to avoid bias by making the measuring process as objective as possible. Despite this there was still some subjectivity in choosing the exact start and endpoints of the measurements, particularly for embryos where the image was less in focus or the view was compromised due to factors such as extracellular fluff, vacuoles, or overlapping PN.

A couple of features that we investigated were subject to dispute; zygote ZP thickness variation and vacuoles. We found no significant difference in zygote ZP thickness variation between any groups. This agrees with the findings of Lewis et al.[31] who found ZP thickness variation was not related to implantation, however it contrasts the findings of Cohen et al[32] who found it was linked to implantation and Sun et al[30] who found higher variation in ZP thickness was associated with higher chance of clinical pregnancy. It is possible that the differences in findings here are due to differing methodologies of calculating thickness variation. Sun et. al and Cohen et al. did not pre-specify the locations where the ZP should be measured, whereas we only took measurements at fixed points. It is possible that our method led to the thinnest or thickest part of the ZP often not being sampled, giving a less representative measure of thickness variation. However, our method also reduces potential introduction of bias between groups, which may also explain the differing findings.

We found a significant difference in vacuoles at zygote stage for both LB vs NP and LB vs discarded. A study by Ebner et al.[16] agreed with these findings, reporting that the presence of vacuoles was inversely correlated with blastocyst formation. Conversely, a study by Barberet et al.[10] reported that there was no link between vacuoles at zygote stage and transfer outcome. It is possible that these results differ from ours due to their relatively small sample size (232 embryos) and/or the fact that they only included high-grade embryos that had been transferred. It is possible that vacuoles at zygote stage are inversely correlated with the chance of forming a blastocyst but do not reduce chance of successful transfer in embryos that have successfully formed blastocysts. As our study included all embryos rather than just those that were highly graded our NP and discarded groups contain more embryos that do not form full blastocysts than our LB group, which could explain the difference between our results and those of Barberet et al. Another possible explanation for differences between our results and published literature (for both vacuolisation and ZP thickness variation) is that there may be differences in the patient population between clinics which are reflected in embryo morphology. Certain morphological markers indicating lower embryo viability could be linked to certain causes of infertility that may be more of less prevalent in different patient populations.

In addition to vacuoles at zygote stage, three other features showed significant difference between the NP and LB groups: morula compaction (perimeter method), ZP thickness at blastocyst stage, and RP. To our knowledge the link between these features and transfer outcome has not been studied before, although ZP thinning is an established component of embryo grading which has been linked to embryo viability[42]. Although no other features were significant, many more were significant when comparing LB to discarded, and the average value of the NP group generally fell in between. This suggests that these features do indicate poor embryo quality, however occur too infrequently in the NP group to be significant.

Investigations into our final datasets via dimensionality reduction and unsupervised clustering led to a greater understanding of the structure in the data. Both the PCA and t-SNE visualisations and the K-means clustering results suggested the presence of a large, slightly more successful than chance cluster and multiple smaller, less successful clusters. The presence of the large cluster, composed of a mix of NP and LB embryos, but with a higher than chance proportion of LB embryos, could suggest that there is a ‘standard’ appearance indicating good quality embryos. The lack of any sizeable higher success rate clusters suggests that many of the NP embryos are indistinguishable from the LB embryos.

Another interesting insight provided by the clustering is that as *k* was increased beyond 2 there appeared to be more ‘unsuccessful’ clusters than ‘successful’ clusters. This suggests that viable embryos are more morphologically uniform than unviable embryos and could mean that there are various different developmental issues causing embryo failure rather than a single or very few common issues. This hypothesis is also supported by the fact that although all our markers were more common in the NP than LB group, they still all occurred quite rarely. The t-SNE visualisation also suggested that there may be certain types of embryos that are very unlikely to lead to live birth, as clusters were seen that had no LB embryos.

Finally, we used a variety of supervised learning models with differing pre-processing options to predict live birth. We chose commonly used and easy to understand models; linear regression, logistic regression, decision tree and support vector machine. Pre-processing with PCA appeared to slightly improve performance by reducing overfitting. The model that achieved the highest test set ROC AUC was linear regression, it is not clear why this model performed best, but it is unlikely to be due solely to reduced overfitting as it also performed best on the training data. We also found that we got better results using the non-binary dataset. This suggests that it is possible for the continuous values of the CNN model outputs to be useful for supervised learning prediction. However, the threshold used for converting the data to binary was chosen arbitrarily, and it is possible that further experimentation with possible threshold values might retain the necessary information whilst reducing noise.

Other studies have investigated the ability of supervised ML algorithms to combine a variety of morphological and/or clinical features to predict embryo viability[23,44-47], however as far as we are aware our study is the first to combine CNN outputs with annotated features to predict live birth using only information from cleavage stage and earlier. Our best model and pre-processing combination achieved a ROC AUC of up to 0.64 on the test dataset. This is a noticeable improvement from the highest performance achieved on our data by any of the previously trained DL models individually, which was a test set ROC AUC of 0.61 for the PN model. Our model performs considerably above chance and is approaching the limit of what is thought to be possible to achieve with just morphokinetic data[45]. This provides a strong argument that selection at earlier stages could be feasible, giving clinics the option of avoiding the potential adverse health outcomes of blastocyst culture. It is important to note that our modelling system first needs to be tested on a hold-out blinded test dataset and data from other clinics. Nevertheless, this study has demonstrated exciting potential for earlier embryo assessment, with the possibility of even better assessment in the future if additional morphological quality markers were to be discovered and included.

Conclusion

We have investigated various potential viability markers for pre-implantation embryos and identified features that can be combined with deep learning predictions to give an improved prediction of live birth pre-blastocyst stage.

Methods

Data

All timelapse videos were provided by the IVF clinic in the Department of Reproductive Medicine at Old Saint Mary's hospital, Manchester University NHS Foundation Trust, Manchester, UK. The developing embryos were cultured in GTL overlaid with Ovoil (Vitrolife, Sweden) from post ICSI (around 40 hours after hCG trigger) to developmental day 5. The media was not changed during the culture period and the embryos were not removed for observation. There were no changes to culture conditions, including culture media throughout the study period. All embryos were cultured and recorded in either the Embryoscope ^{TM} or the Embryoscope+ ^{TM} (Vitrolife, Sweden) at 37°C, 6% CO_2 and 5% O_2. Each time-lapse video exported was 500x500 pixels and had a framerate of 5-10 frames an hour. All videos were irreversibly anonymised by staff at the clinic before we received them.

The dataset included ICSI embryos from 2016-2019 with three different outcomes; live birth, no pregnancy, or discarded due to being considered too low quality to transfer. We included all single embryo transfers (SET) and double embryo transfers (DET) that resulted in either no pregnancy or different gender twins.

Quantitative measurements

The time at 3 cell measurement was carried out using the timestamp and all other measurements were carried out using Imagej. All PN and zona measurements were carried out using the 'straight line' tool and were taken 1 hour before NEBD at zygote stage and at the last frame before embryo was taken out for transfer at blastocyst stage.

The morula stage measurements were carried out at the frame where the embryo was judged to be most compact. Two techniques were used; angles and perimeter. The angles technique used the angle tool on Imagej to measure all angles between outside blastomeres (see Fig. S1a for illustration) that could be seen in that plane. This resulted in 1-8 angles being recorded for each embryo, all of which were used to calculate the 'average angle'. The perimeter technique used the freehand selection tool to trace the perimeter of the embryo (Fig. S1a) and the oval tool to draw a bounding circle round the embryo. The magnitude of the difference between the circle circumference and the perimeter was then calculated for each embryo.

Qualitative feature annotation

All qualitative features, or 'markers' not occurring at a specific moment (e.g 'cleavage furrow from one side') were all assessed at specific times before or after a landmark moment as this allows for quicker classification (so more likely to be adopted by embryologists in practice) and helps to reduce subjectivity. Cytoplasmic fragments were assessed 7 hours after FD as cells have stopped rolling around but not started dividing yet. rough patch (RP) and zygote vacuoles were assessed 0.5 hours after last sighting of the PN is this it is as close as possible to a landmark event and does not have the PN obscuring the view of some of the cytoplasm. Asymmetric PN was assessed 0.5 hours before last trace because we observed that it is common for the PN to move into a central position just before NEBD even in healthy embryos.

Dimensionality reduction algorithms

Dimensionality reduction was carried out using the scikit-learn PCA and t-SNE function. Features were first standardised by removing the mean and scaling to unit variance using the scikit-learn StandardScaler function. Default parameters were used for both PCA and t-SNE, with the exception of varying the number of components for PCA and the perplexity for t-SNE, and also using the 'jaccard' distance metric for t-SNE in the binary dataset (as it is suitable for binary data).

Modelling

For supervised learning four different models were used, all with sci-kit-learn functions; LinearRegression, LogisticRegression, tree, and SVC. Default parameters were used for all models apart from class weight which was always set to balance out the class size inequality (apart from in linear regression where this was not an option) and min-leaf in the decision tree, which was iteratively varied from 1 to 100. The maximum test score found over all values of min-leaf was reported, which used a value of 75. Each model was trained 50 times with a different random train/test split, with 50 embryos from the LB group and 96 embryos from the NP group in the test set and the remaining embryos all in the training set. The ROC AUC was calculated for each model training attempt using the scikit-learn function roc_auc_score. An average ROC AUC was then calculated for each model over the 50 runs. For unsupervised learning we used the KMeans scikit-learn function with default parameters apart from cluster number which was varied from 2 to 15.

Graph production

All box plots were produced using the Seaborn library in python and all clustered column plots were produced in Microsoft Excel. All scatter plots were produced in the using matplotlib.pyplot in python. For the clustering graph (Fig.3b) the size of each point was set by calculating the size of the cluster it represented as a fraction of the whole dataset and then multiplying by 100 so the points were easily visible. For the non-binary features PCA and t-SNE scatter plots (Fig. 2b) an alpha value of 0.2 and a jitter value of 0.02 (in both x and y) were used to make the points partially transparent and randomly offset so that overlapping data-points could be seen.

Statistics

All continuous data was first tested for normality using the D'Agostino & Pearson test. For cases where two data sets were compared and the data was normally distributed, a students t-test was used to test for significance. If the data was not normally distributed, a Mann Whitney test was performed. For cases where there were more than two datasets being compared and data was normally distributed a one-way ANOVA test was used, followed by a post-hoc test (HSD) if the result was significant. For cases where there were more than two datasets being compared and data was not normally distributed a Kruksall-Wallis test was used, followed by a post-hoc test (Dunns) if the result was significant. For the categorical data chi-squared tests were used to calculate significance. For all significance tests, results were said to be significant if $p<0.05$. Where significance is shown on graphs, * corresponds to $p<0.05$, ** corresponds to $p<0.01$, and *** corresponds to $p<0.001$

The errors reported for ROC AUC of the supervised learning algorithms were the standard error of the mean calculated from the ROC AUC values of all 50 training iterations for each model.

References

1 Vander Borght, M. & Wyns, C. Fertility and infertility: Definition and epidemiology. *Clinical biochemistry* **62**, 2-10 (2018). https://doi.org:10.1016/j.clinbiochem.2018.03.012

2 Steptoe, P. C. & Edwards, R. G. BIRTH AFTER THE REIMPLANTATION OF A HUMAN EMBRYO. *The Lancet (British edition)* **312**, 366-366 (1978). https://doi.org:10.1016/S0140-6736(78)92957-4

3 Thurin, A. *et al.* Elective Single-Embryo Transfer versus Double-Embryo Transfer in in Vitro Fertilization. *The New England journal of medicine* **351**, 2392-2402 (2004). https://doi.org:10.1056/NEJMoa041032

4 Martikainen, H., Orava, M., Lakkakorpi, J. & Tuomivaara, L. Day 2 elective single embryo transfer in clinical practice: better outcome in ICSI cycles. *Human reproduction (Oxford)* **19**, 1364-1366 (2004). https://doi.org:10.1093/humrep/deh197

5 Le Lannou, D. *et al.* Contribution of embryo cryopreservation to elective single embryo transfer in IVF–ICSI. *Reproductive biomedicine online* **13**, 368-375 (2006). https://doi.org:10.1016/S1472-6483(10)61441-1

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

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

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

9 Cruz, M., Muñoz, M. & Meseguer, M. 45-53 (Springer New York, 2013).

10 Barberet, J. *et al.* Can novel early non-invasive biomarkers of embryo quality be identified with time-lapse imaging to predict live birth? *Human reproduction (Oxford)* **34**, 1439-1449 (2019). https://doi.org:10.1093/humrep/dez085

11 Otsuki, J. *et al.* Noninvasive embryo selection: kinetic analysis of female and male pronuclear development to predict embryo quality and potential to produce live birth. *Fertility and sterility* **112**, 874-881 (2019). https://doi.org:10.1016/j.fertnstert.2019.07.015

12 Manor, D., Drugan, A., Stein, D., Pillar, M. & Itskovitz-Eldor, J. Unequal pronuclear size : A powerful predictor of embryonic chromosome anomalies. *Journal of assisted reproduction and genetics* **16**, 385-389 (1999). https://doi.org:10.1023/A:1020550115345

13 Nagy, Z. P. *et al.* Pronuclear morphology evaluation with subsequent evaluation of embryo morphology significantly increases implantation rates. *Fertility and sterility* **80**, 67-74 (2003). https://doi.org:10.1016/S0015-0282(03)00569-7

14 Sadowy, S., Tomkin, G., Munné, S., Ferrara-Congedo, T. & Cohen, J. Impaired development of zygotes with uneven pronuclear size. *Zygote (Cambridge)* **6**, 137-141 (1998). https://doi.org:10.1017/S0967199498000057

15 Nasiri, N. & Eftekhari-Yazdi, P. An overview of the available methods for morphological scoring of pre-Implantation embryos in in vitro fertilization. *Cell journal (Yakhteh)* **16**, 392-405 (2015).

16 Ebner, T. *et al.* Occurrence and developmental consequences of vacuoles throughout preimplantation development. *Fertility and sterility* **83**, 1635-1640 (2005). https://doi.org:10.1016/j.fertnstert.2005.02.009

17 Ebner, T. *et al.* Presence, but not type or degree of extension, of a cytoplasmic halo has a significant influence on preimplantation development and implantation behaviour. *Human reproduction (Oxford)* **18**, 2406-2412 (2003). https://doi.org:10.1093/humrep/deg452

18 Prados, F. J., Debrock, S., Lemmen, J. G. & Agerholm, I. The cleavage stage embryo. *Human reproduction (Oxford)* **27**, i50-i71 (2012). https://doi.org:10.1093/humrep/des224

19 Pera, R. A. R. *et al.* Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nature biotechnology* **28**, 1115-1121 (2010). https://doi.org:10.1038/nbt.1686

20 Hardarson, T., Hanson, C., Sjögren, A. & Lundin, K. Human embryos with unevenly sized blastomeres have lower pregnancy and implantation rates: indications for aneuploidy and multinucleation. *Human reproduction (Oxford)* **16**, 313-318 (2001). https://doi.org:10.1093/humrep/16.2.313

21 Van Royen, E. *et al.* Calculating the implantation potential of day 3 embryos in women younger than 38 years of age: a new model. *Human reproduction (Oxford)* **16**, 326-332 (2001). https://doi.org:10.1093/humrep/16.2.326

22 Scott, L., Finn, A., O'Leary, T., McLellan, S. & Hill, J. Morphologic parameters of early cleavage-stage embryos that correlate with fetal development and delivery: prospective and applied data for increased pregnancy rates. *Human reproduction (Oxford)* **22**, 230-240 (2007). https://doi.org:10.1093/humrep/del358

23 Basile, N. *et al.* The use of morphokinetics as a predictor of implantation: a multicentric study to define and validate an algorithm for embryo selection. *Human reproduction (Oxford)* **30**, 276-283 (2015). https://doi.org:10.1093/humrep/deu331

24 Iwata, K. *et al.* Analysis of compaction initiation in human embryos by using time-lapse cinematography. *Journal of assisted reproduction and genetics* **31**, 421-426 (2014). https://doi.org:10.1007/s10815-014-0195-2

25 Cruz, M. *et al.* Timing of cell division in human cleavage-stage embryos is linked with blastocyst formation and quality. *Reproductive biomedicine online* **25**, 371-381 (2012). https://doi.org:10.1016/j.rbmo.2012.06.017

26 Cetinkaya, M. *et al.* Relative kinetic expressions defining cleavage synchronicity are better predictors of blastocyst formation and quality than absolute time points. *Journal of assisted reproduction and genetics* **32**, 27-35 (2015). https://doi.org:10.1007/s10815-014-0341-x

27 Basile, N. P. D. *et al.* Increasing the probability of selecting chromosomally normal embryos by time-lapse morphokinetics analysis. *Fertility and sterility* **101**, 699-704.e691 (2014). https://doi.org:10.1016/j.fertnstert.2013.12.005

28 Campbell, A. *et al.* Retrospective analysis of outcomes after IVF using an aneuploidy risk model derived from time-lapse imaging without PGS. *Reproductive biomedicine online* **27**, 140-146 (2013). https://doi.org:10.1016/j.rbmo.2013.04.013

29 Lemmen, J. G., Agerholm, I. & Ziebe, S. Kinetic markers of human embryo quality using time-lapse recordings of IVF/ICSI-fertilized oocytes. *Reproductive biomedicine online* **17**, 385-391 (2008). https://doi.org:10.1016/S1472-6483(10)60222-2

30 Sun, Y. P., Xu, Y., Cao, T., Su, Y. C. & Guo, Y. H. Zona pellucida thickness and clinical pregnancy outcome following in vitro fertilization. *Int J Gynaecol Obstet* **89**, 258-262 (2005). https://doi.org:10.1016/j.ijgo.2005.02.012

31 Lewis, E. I. *et al.* Use of imaging software for assessment of the associations among zona pellucida thickness variation, assisted hatching, and implantation of day 3 embryos. *Journal of assisted reproduction and genetics* **34**, 1261-1269 (2017). https://doi.org:10.1007/s10815-017-0978-3

32 Cohen, J., Wiemer, K. E. & Wright, G. Prognostic value of morphologic characteristics of cryopreserved embryos: a study using videocinematography. *Fertility and sterility* **49**, 827-834 (1988). https://doi.org:10.1016/S0015-0282(16)59892-6

33 Glatstein, I., Chavez-Badiola, A. & Curchoe, C. L. New frontiers in embryo selection. *Journal of assisted reproduction and genetics* **40**, 223-234 (2023). https://doi.org:10.1007/s10815-022-02708-5

34 Fernandez, E. I. *et al.* Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. *Journal of assisted reproduction and genetics* **37**, 2359-2376 (2020). https://doi.org:10.1007/s10815-020-01881-9

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

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

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

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

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

40 Mapstone, C., Hunter, H., Brison, D., Handl, J. & Plusa, B. Deep learning pipeline reveals key moments in human embryonic development predictive of live birth in IVF (University of Manchester, 2023).

41 Maître, J.-L., Niwayama, R., Turlier, H., Nédélec, F. & Hiiragi, T. Pulsatile cell-autonomous contractility drives compaction in the mouse embryo. *Nature cell biology* **17**, 849-855 (2015). https://doi.org:10.1038/ncb3185

42 Zhao, Y.-Y., Yu, Y. & Zhang, X.-W. Overall Blastocyst Quality, Trophectoderm Grade, and Inner Cell Mass Grade Predict Pregnancy Outcome in Euploid Blastocyst Transfer Cycles. *Chinese medical journal* **131**, 1261-1267 (2018). https://doi.org:10.4103/0366-6999.232808

43 Leonavicius, K. *et al.* Mechanics of mouse blastocyst hatching revealed by a hydrogel-based microdeformation assay. *Proceedings of the National Academy of Sciences - PNAS* **115**, 10375-10380 (2018). https://doi.org:10.1073/pnas.1719930115

44 Bodri, D. *et al.* Predicting live birth by combining cleavage and blastocyst-stage time-lapse variables using a hierarchical and a data mining-based statistical model. *Reprod Biol* **18**, 355-360 (2018). https://doi.org:10.1016/j.repbio.2018.10.006

45 Milewski, R., Kuczyńska, A., Stankiewicz, B. & Kuczyński, W. How much information about embryo implantation potential is included in morphokinetic data? A prediction model based on artificial neural networks and principal component analysis. *Advances in medical sciences* **62**, 202-206 (2017). https://doi.org:10.1016/j.advms.2017.02.001

46 De Gheselle, S. *et al.* Machine learning for prediction of euploidy in human embryos: in search of the best-performing model and predictive features. *Fertility and sterility* **117**, 738-746 (2022). https://doi.org:10.1016/j.fertnstert.2021.11.029

47 Uyar, A., Bener, A. & Ciray, H. N. Predictive Modeling of Implantation Outcome in an In Vitro Fertilization Setting: An Application of Machine Learning Methods. *Medical decision making* **35**, 714-725 (2015). https://doi.org:10.1177/0272989X14535984

Supplementary Materials

Figures

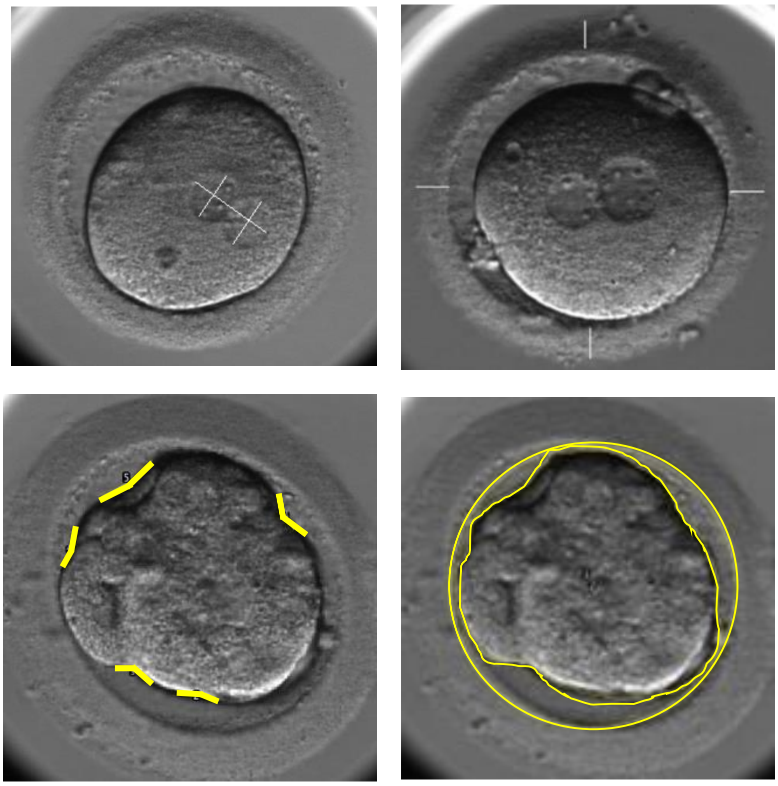


Figure S1a: Illustration of quantitative feature measurement methods. Top left: PN measurements, top right: ZP measurements, bottom left: morula compaction measurement using ‘angle method’, bottom right: morula compaction measurement using ‘perimeter method’.

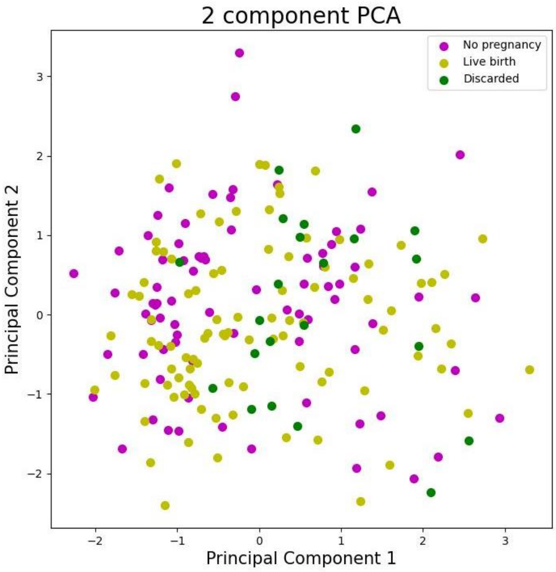


Figure S1b: 2 component PCA of zygote stage features. Including ZP thickness, ZP variation, PN average area, PN area difference.

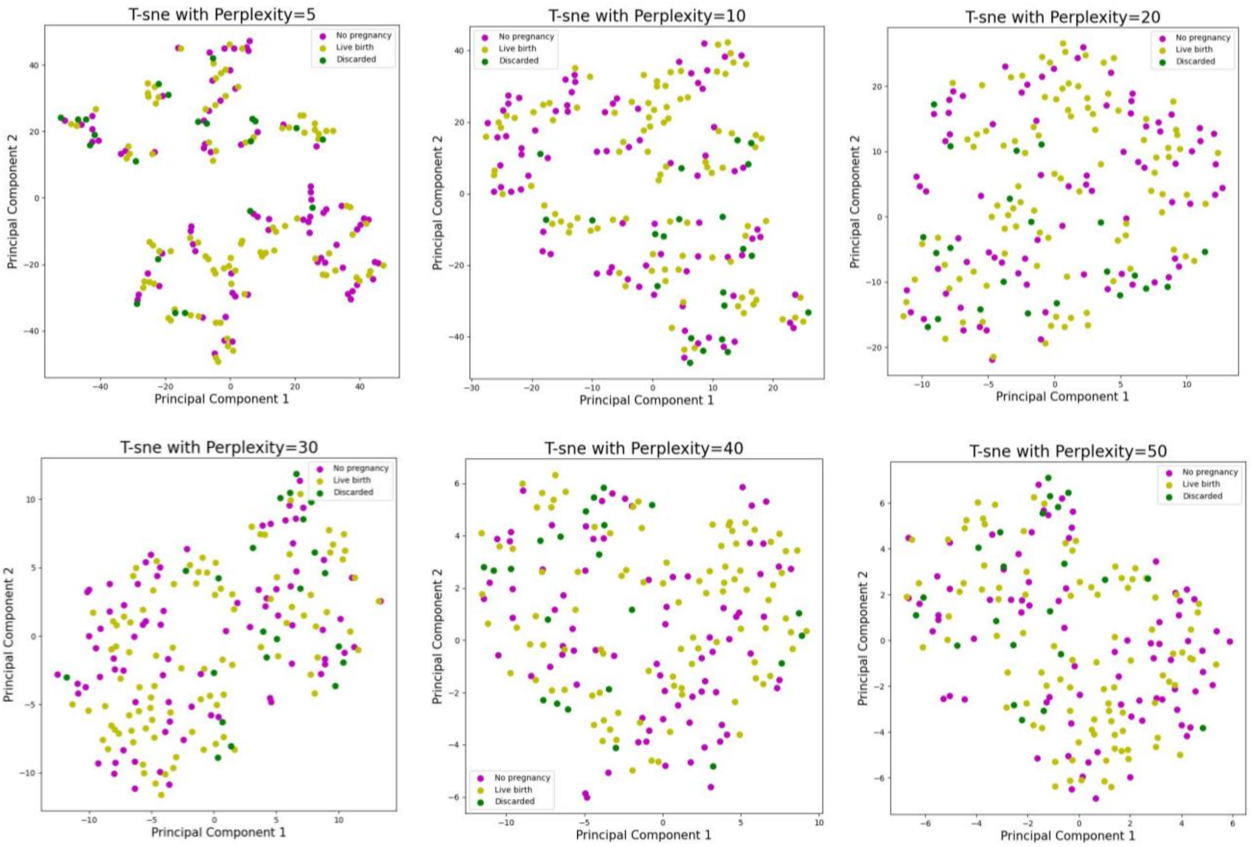


Figure S1c: t-SNE visualisations of zygote stage features with varying perplexity parameters. Including ZP thickness, ZP variation, PN average area, and PN area difference.

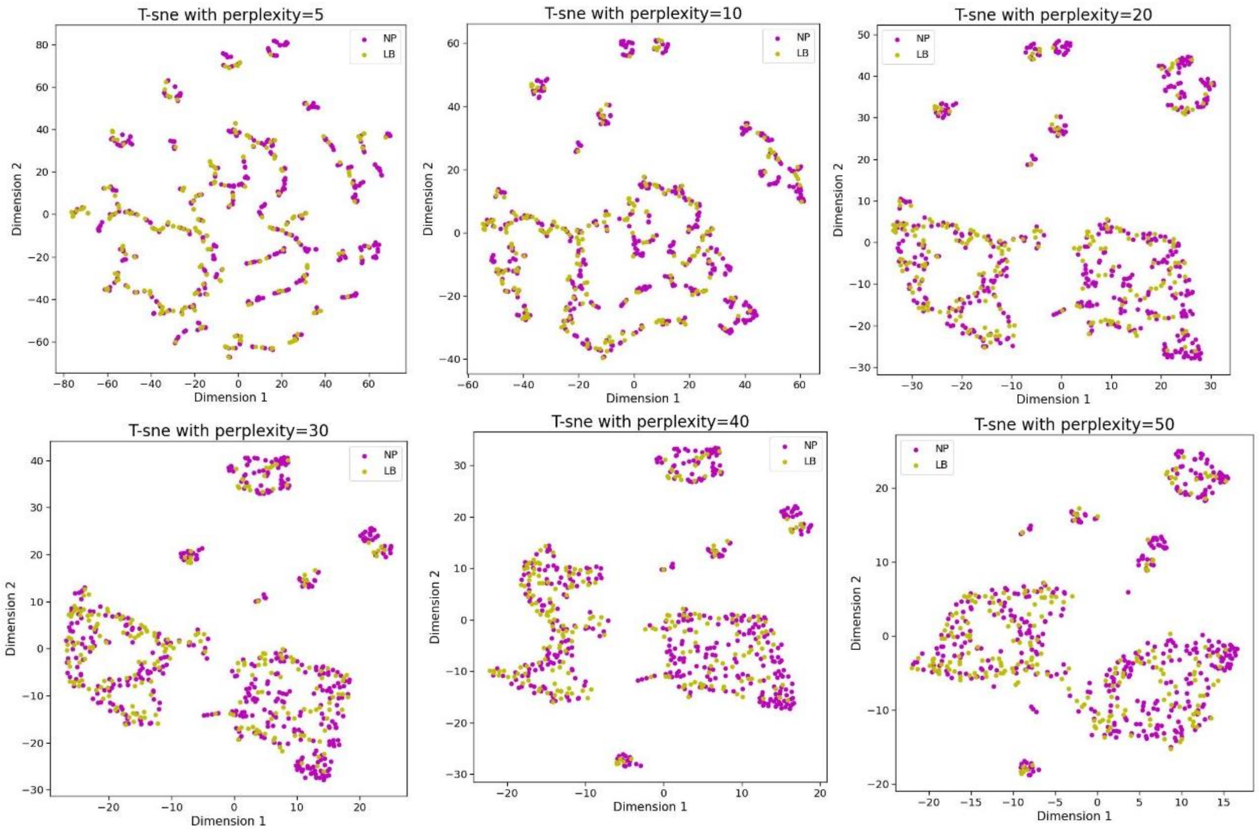


Figure S2a: t-SNE on final dataset with non-binary input features

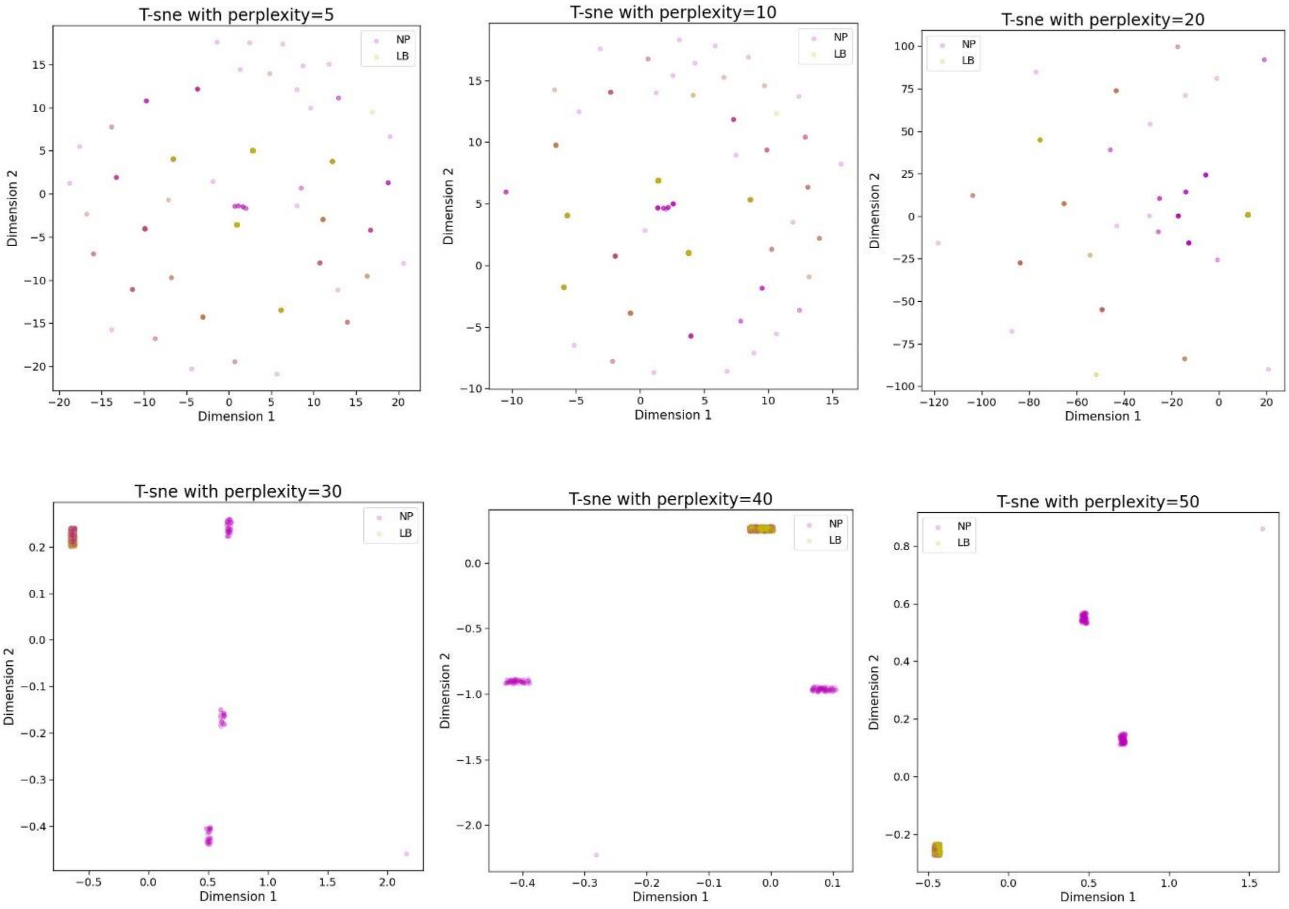


Figure S2b: t-SNE on final dataset with binary input features

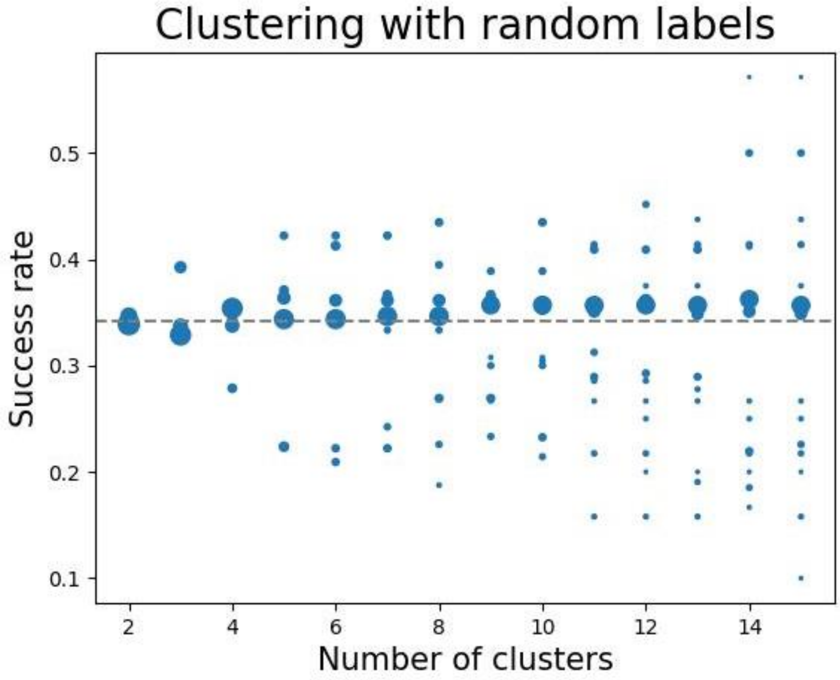


Figure S3a: control experiment for K-means clustering

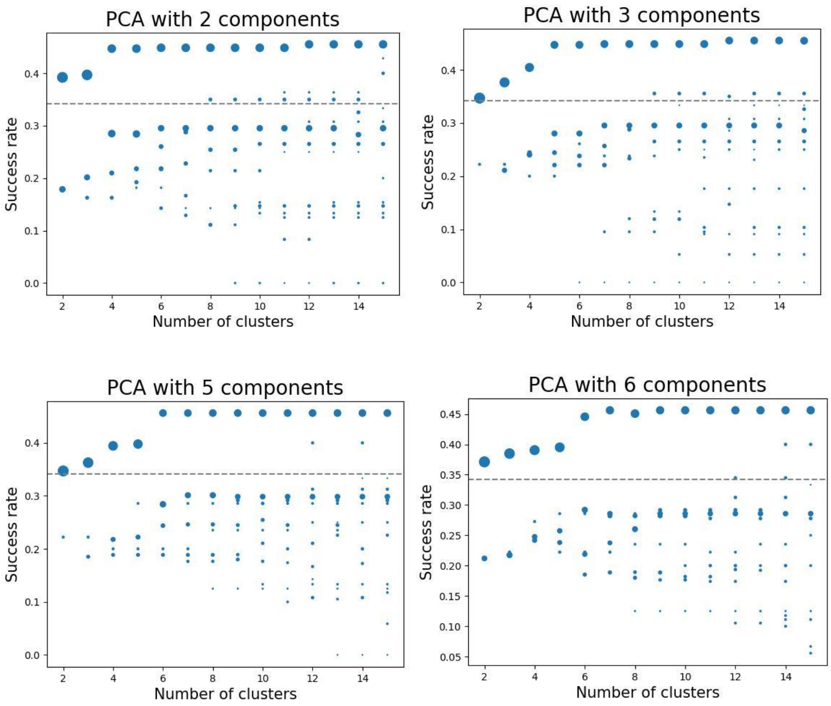


Figure S3b: Clustering results after PCA pre-processing

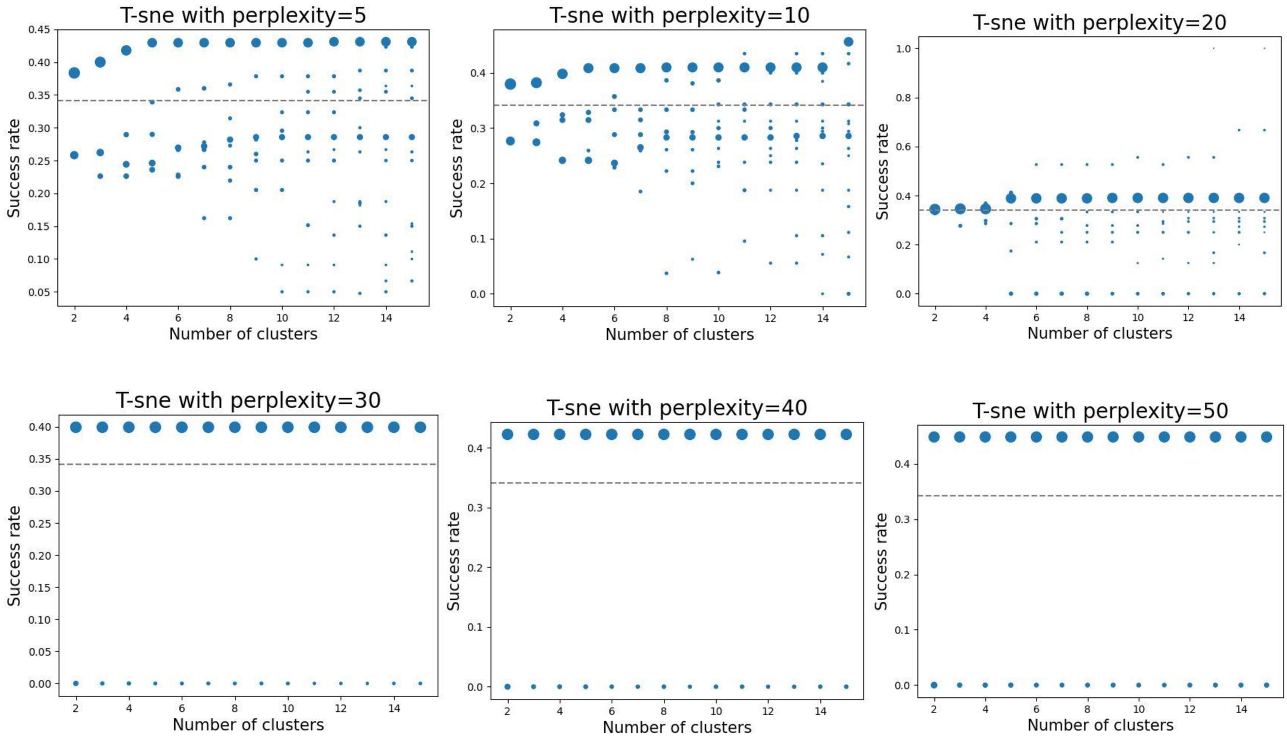


Figure S3c: Clustering results after t-SNE pre-processing, perplexity varying

Tables

No pre-processing non binary		
Modelling system	**Training ROC AUC**	**ROC AUC**
Logistic regression	0.601	0.589 +/-0.005
Linear regression	0.652	0.629 +/- 0.005
Decision tree	0.619	0.553 +/- 0.005
SVC	0.642	0.592 +/- 0.006
No pre-processing binary		
Modelling system	**Training ROC AUC**	**ROC AUC**
Logistic regression	0.609	0.599+/-0.005
Linear regression	0.642	0.624+/-0.005
Decision tree	0.591	0.589+/- 0.004
SVC	0.622	0.582+/- 0.004
PCA non binary 2 components		
Modelling system	**Training ROC AUC**	**ROC AUC**
Logistic regression	0.607	0.603+/- 0.004
Linear regression	0.639	0.635+/- 0.004
Decision tree	0.624	0.590+/- 0.005
SVC	0.597	0.611+/- 0.004
PCA non binary 3 components		
Modelling system	**Training ROC AUC**	**ROC AUC**
Logistic regression	0.604	0.586 +/- 0.004
Linear regression	0.642	0.639 +/- 0.006
Decision tree	0.625	0.578 +/- 0.005
SVC	0.601	0.594 +/- 0.006
PCA non binary 4 components		
Modelling system	**Training ROC AUC**	**ROC AUC**
Logistic regression	0.602	0.594+/- 0.006
Linear regression	0.6428	0.633+/- 0.004
Decision tree	0.626	0.571+/- 0.005
SVC	0.635	0.598+/- 0.005
PCA binary 2 components		
Modelling system	**Training ROC AUC**	**ROC AUC**
Logistic regression	0.601	0.597+/- 0.006
Linear regression	0.6299	0.622+/- 0.004
Decision tree	0.608	0.599+/- 0.004
SVC	0.604	0.585+/- 0.005
PCA binary 3 components		
Modelling system	**Training ROC AUC**	**ROC AUC**
Logistic regression	0.600	0.597+/- 0.004
Linear regression	0.631	0.627+/- 0.004
Decision tree	0.609	0.601+/- 0.005
SVC	0.604	0.578+/- 0.005
PCA binary 4 components		

Modelling system	Training ROC AUC	ROC AUC
Logistic regression	0.602	0.602+/- 0.004
Linear regression	0.631	0.624+/- 0.004
Decision tree	0.609	0.590+/- 0.005
SVC	0.607	0.5845+/- 0.006

Table S1: supervised learning results. The training and test set ROC AUC for live birth prediction is shown for a variety of model types and pre-processing methods.

Chapter 4:

Nuclei dimensions shown to correlate with developmental stage and cell lineage in early mouse embryos

Camilla Mapstone, Katarzyna Filimonow, Jessica Forsyth, Isobel Green, Adepeju Adedeji, Daniel Brison, Julia Handl, Berenika Plusa

Rationale of paper

The research presented in chapter 3 demonstrated that sub-cellular features from the zygote stage could be useful in live birth prediction, however there is currently a lack of understanding of how sub-cellular features, such as the appearance of the nucleus, might be related to embryo viability from cleavage stage onwards. First, an understanding of the typical pattern of morphological changes is needed so any deviations from normal development can be detected. Therefore, the paper presented in this chapter investigates changes in nuclear size and shape over developmental stage and cell lineage. For this, mouse embryos were used due to ease of accessibility and high quality compared to human embryos. This paper is a draft manuscript that we intend to submit later this year. All supplementary material is included at the end of the manuscript.

Aims

-Investigate how mouse nuclei size and shape vary over development from 8 cell stage to E5.5.

-Investigate how mouse nuclei size and shape vary across cell lineage at the same developmental stage.

Author contributions

The original conceptualisation of the research questions addressed in this work was developed by Dr. Berenika Plusa. Data collection was performed by myself, Dr. Berenika Plusa and Dr Katarzyna Filimonow. Data processing was carried out by myself, Dr Katarzyna Filimonow and two Masters students I helped supervise (Isobel green and Adepeju Adedeji). All subsequent analyses and figure creation were then completed by me. I designed and performed the unsupervised clustering experiments and analysed the results. The manuscript was written by myself and critically evaluated by Dr. Berenika Plusa and Dr. Julia Handl.

Nuclei dimensions shown to correlate with developmental stage and cell lineage in the early mouse embryo

Camilla Mapstone[1,2], Katarzyna Filimonow[3], Jessica Forsyth[4], Isobel green[1] Adepeju Adedeji[1], Julia Handl[2], Berenika Plusa[1]

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

[2]Alliance Manchester Business School, University of Manchester, Manchester, UK

[3]Department of Embryology, Faculty of Biology, The University of Warsaw, I. Miecznikowa, Warsaw, Poland

[4]School of Mathematics, Alan Turing Building, University of Manchester, Manchester, UK

Abstract

Nuclear shape and size is thought to be an important factor of healthy development, however there is still little known about the typical development of nuclear morphology over early mammalian embryonic stages. Here, we have investigated changes in the shape and size of mouse embryo nuclei from 8 cell stage to E5.5. We found that size generally decreases until the 128 cell stage, before increasing at E5.5, and the nuclei also follow a general trend of increasing ellipticity over this period. We also found that there were significant differences between nuclei of differing lineages from the 32 cell stage onwards, opening the possibility to use nuclear shape and size for non- invasive assessment of lineage formation in mammalian embryos.

Introduction

Mammalian embryonic development is a self-organised process in which the very first cell lineages are established. First to appear is the trophectoderm (TE) and inner cell mass (ICM). The TE takes an outside position and eventually contributes to the embryonic part of the placenta[1,2], while the ICM contains the inner group of cells that are pushed to one side of the embryo by the growing cavity[3]. Next, the ICM cells further differentiate into the epiblast (EPI) which eventually goes on to form the fetus, and the extra-embryonic primitive endoderm (PrE)[4]. Following implantation, the TE further differentiates into the placenta progenitors; the extraembryonic ectoderm (ExE) and the ectoplacental cone, while the PrE forms the parietal endoderm and the visceral endoderm (VE), which gives rise to the endoderm of the visceral yolk sac[5]. Despite much progress in the last few decades, the processes that give rise to these first few lineages are still not fully understood[6]

A greater understanding of the sub-cellular changes accompanying pre-implantation development could be very beneficial to the field of IVF, the progress of which is currently restricted due to the limited amount of knowledge available to inform embryo selection procedures[6-8]. Although nuclei appearance been studied at the zygote stage[9-14], the normal morphology of nuclei at later stages has not yet been investigated, even though nuclei are visible throughout development using the latest time-lapse systems. Knowledge of the regulation of nuclei size and shape is also limited in general. A greater understanding of the mechanisms that control dimensions of the nuclei could have additional important implications for broader medical research as deviations from normal nuclei size and shape are linked to disease[15].

Previous research into changes in nuclei over early embryonic development has revealed that the size of nuclei in mouse embryos decreases from zygote to early blasyocyst[16], however it is not yet known whether this trend continues into the peri-implantation period. Forsyth et al. (2021)[17] showed that the same trend exists in cell diameter with the diameter decreasing up to the E4.5 stage, but then the trend reverses and cell diameter increases from E4.5 to E5.5. It remains to be seen whether nuclei size follows the same pattern. The change in nuclei shape over development has also not yet been studied, however it has previously been shown that cells become less spherical from the 2 cell to 64 cell stage[18]

Our knowledge of changing nuclei dimensions across early development is currently incomplete, in this study we have set out to establish a complete picture of how nuclei size and shape varies throughout this time period in mouse embryos. We first examined the size and shape of nuclei from the 8 cell to E5.5 stage, finding a general trend of decreasing size and roundness as the embryo develops. Next, we investigate how the nuclei dimensions vary across different lineages at the same developmental stage and discovered statistically significant lineage-based differences at all stages with identifiably separate lineages. Finally, we used unsupervised machine learning to further explore the distribution of nuclei shapes in different cell lineages at the 128 cell stage. Our data reveal the existence of lineage specific biases in nuclear shape at the time of implantation.

Results

In this study we have investigated the variation in nuclei size and shape using 3D confocal z-stack fluorescence images of live, disaggregated, and fixed mouse embryos. This involved manually measuring individual nuclei using the Imagej and IMARIS software packages. These measurements were carried out on embryos ranging from the 8 cell stage, where it is believed the lineage specification process first begins in the mouse, to the post-implantation E5.5 stage.

Nuclear changes over development

We first investigated how the size and shape of nuclei change across development. Firstly, we analysed the data from Forsyth et al. (2021)[17] to measure the cell nuclei. We used the data set containing images of embryos from the H2B-GFP strain of mice, therefore measurements could be taken using the H2B-GFP fluorescence. The embryos in this dataset were disaggregated, using calcium-free M2 for embryos up to E3.5 and treatment with trypsin for E4.5 and E5.5 embryos.

The nuclei were measured in the Z-plane with the greatest cross-sectional area. Measurements were only taken in the XY-plane because the resolution was much lower across the Z-plane and 2D measurements were considered to be sufficient as orientation of nuclei is likely to be random when disaggregated cells are placed randomly in a drop of medium. Two measurements were taken for each nucleus, one along the longest possible diameter and a second perpendicular measurement, these can be thought of as major and minor axes measurements if the nuclei shape is assumed to be an ellipse. For each nucleus we then averaged over the two measurements to get the average diameter and also calculated circularity by dividing the smaller diameter by the larger diameter. The results of this are shown in Fig. 1b. The nuclei appear to decrease in size from 8 cell to E4.5, with a possible increase (albeit non-significant) from E4.5 to E5.5. The circularity results showed a general decrease from the 8 and 16 cell stage to the E3.5 to E5.5 stage, however circularity did not appear to differ much between the 8 cell stage and 16 cell stage.

We then wanted to see whether these trends in nuclei variation over development held up in intact embryos. For this we used live images of freshly flushed H2B-GFP mice embryos from 8 cell stage to E5.5. As we were able to count the number of cells in each embryo we grouped embryos by cell size, with the E3.5-E4.5 embryos now divided into 32, 64 or 128 cell groups. We included embryos that had cell counts within 10% of 32, 64, or 128. First, we took 2D measurements of cell nuclei following the same approach as for the disaggregated embryos, the results are shown in Fig. 1c. For average diameter we see the same trend as in disaggregated embryos, diameter decreases from 8 cell to 128 cell and then increases to E5.5, however the increase from 128 to E5.5 is now statistically significant. For circularity there is a similar, but smoother, trend as before, with circularity steadily decreasing from 16 cell stage to E5.5. Again, circularity does not appear to differ between 8 to 16 cell stage.

Next, we repeated the measurements in the intact embryos using 3D approaches. Although this introduces increased subjectivity and measurement error due to lower Z resolution, the nuclei are no longer randomly orientated when held in the structure of the embryo so 3D measurements may be needed to fully describe the dimensions of the nuclei. Two methods were used for the 3D approach; manual measurements and semi-automated surface detection. For the former, three manual measurements were taken for each nuclei; the longest diameter (the 3D equivalent of the major axis of an ellipse) and the diameter along the two axes perpendicular to this. The volume and sphericity were then calculated as described in material and methods section.

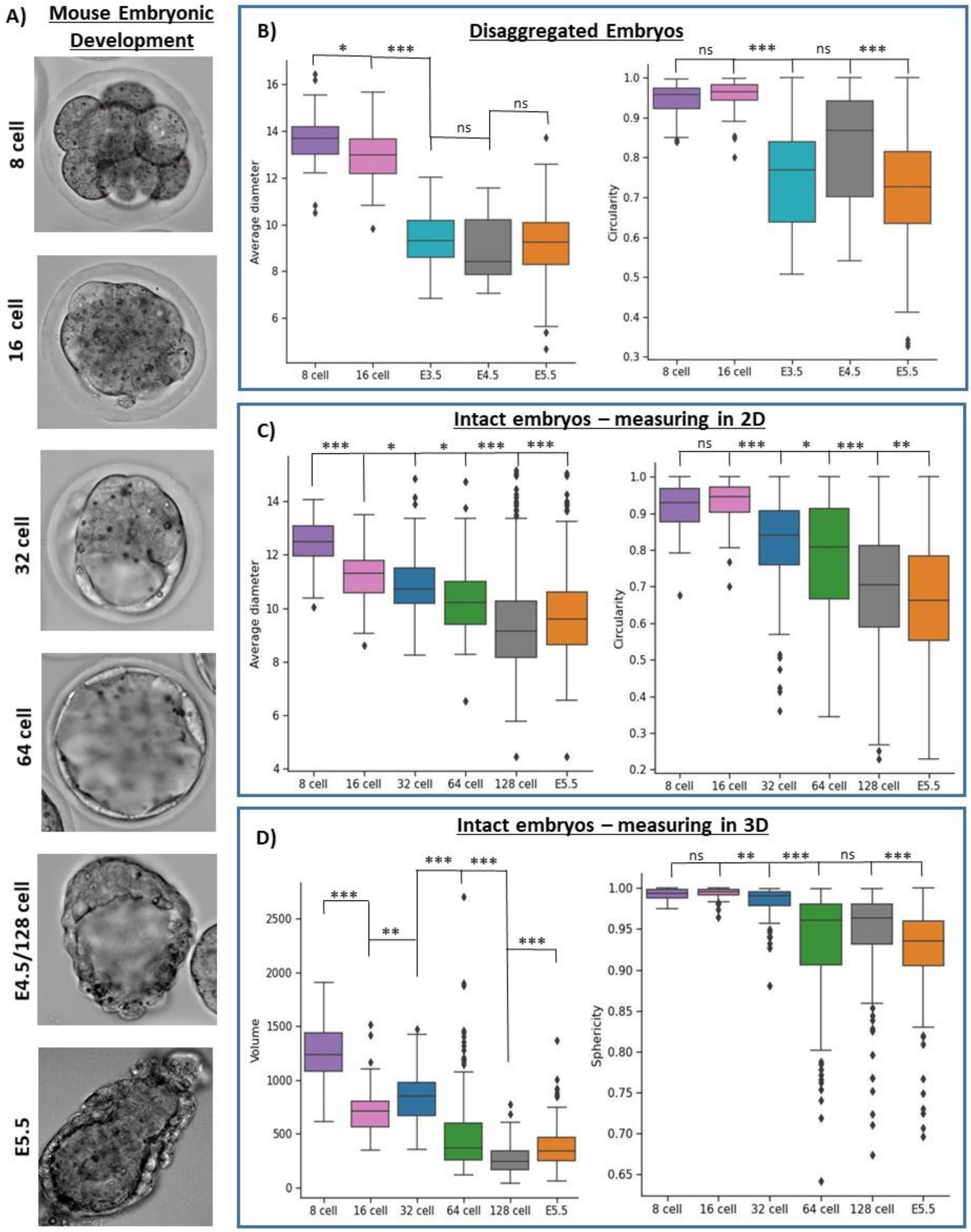


Figure 1: Changes in Nuclei dimensions over development. Left: Pre-implantation and peri-implantation development in the mouse embryo. Bright field images of freshly flushed embryos from 8 cell stage to E5.5 are shown. A) Average diameter and circularity of nuclei in disaggregated embryos with measurements taken in 2D. A total of 25

embryos were included. An overall p value of $1.21\text{x}10^{-58}$ (from Kruksall-Wallis) was found for average diameter and $3.91\text{x}10^{-50}$ (from Kruksall-Wallis) for circularity B) Average diameter and circularity of nuclei in intact embryos with measurements taken in 2D. A total of 26 embryos were included. An overall p value of $2.63\text{x}10^{-83}$ (from Kruksall-Wallis) was found for average diameter and $2.19\text{x}10^{-78}$ (from Kruksall-Wallis) for circularity C) Volume and sphericity of nuclei in intact embryos, measurements taken in 3D. A total of 26 embryos were included. An overall p value of $3.68\text{x}10^{-70}$ (from Kruksall-Wallis) was found for volume and $2.90\text{x}10^{-52}$ (from Kruksall-Wallis) for sphericity. All box plots represent all individual nuclei from all embryos of that developmental stage. Following ANOVA/ Kruksall-Wallis tests, post-hoc tests (HSD/Dunns) were performed to make sequential pairwise comparisons.

The results for volume and sphericity across developmental stage are shown in Fig. 1d. The volume appears to follow the same general trend as seen for average diameter (with the exception of the 16 to 32 cell stage transition) where nuclear volume generally decreases from 8 cell to 128 cell before rising significantly from 128 cell to E5.5. In general, sphericity decreases over development, as observed for circularity, with the exception of the 64 to 128 transition, where no statistically significant change in sphericity was observed.

We also attempted an automated approach to measuring the nuclei in 3D using the IMARIS surfaces tool, results are shown in Fig S1. However, we found that there was a lot of subjectivity in the post-processing correction steps and the tool became very difficult to use at later stages due to nuclei being closer together and the tool not allowing for any overlap between surfaces. Nevertheless, the results still showed a general downwards trend when looking at both volume and sphericity over development.

Overall, our results suggest that nuclei size decreases from the 8 cell to 128 cell stage, before increasing from 128 cell stage to E5.5, while the circularity/sphericity generally decreases from the 16 cell stage onwards. We have observed the same trends in intact and disaggregated embryos, regardless of the method used.

Nuclear dimension varies by lineage

Upon discovering that nuclei size and shape changes throughout development, we next wanted to investigate how nuclei size and shape vary across lineages at the same developmental stage. For this we used embryos that had been fixed and immunostained for 8 to 128 cell stage and continued to use the live embryos for E5.5 as the lineages are clearly distinguishable in brightfield at this stage (see Fig. 2). We stained embryos with SOX2 and GATA4, SOX2 is an ICM marker at 32 cell stage and a EPI marker from 64 cell stage onwards, while GATA4 is a PrE marker. We found that SOX2 was present in some cells from the 32 cell stage onwards and GATA4 was also present in some cells from the 64 cell stage onwards. Nuclei that were negative for both SOX2 and GATA4 were categorised as TE for the 32-128 cell stage (due to their position in the embryo and lack of staining for EPI and PrE markers). Measurements were taken using the same approach as before, but using the fluorescence from the nuclear Hoechst staining rather than H2B-GFP for the fixed embryos.

Results for each stage are shown in Fig. 2. At the 32 cell stage the SOX2 negative TE nuclei are significantly less circular and spherical than the SOX2 positive nuclei. The TE nuclei also appear to be bigger than the SOX2 positive nuclei, although this is only significant in 2D. These results suggest that TE nuclei are more elliptical and slightly bigger than ICM nuclei at this stage.

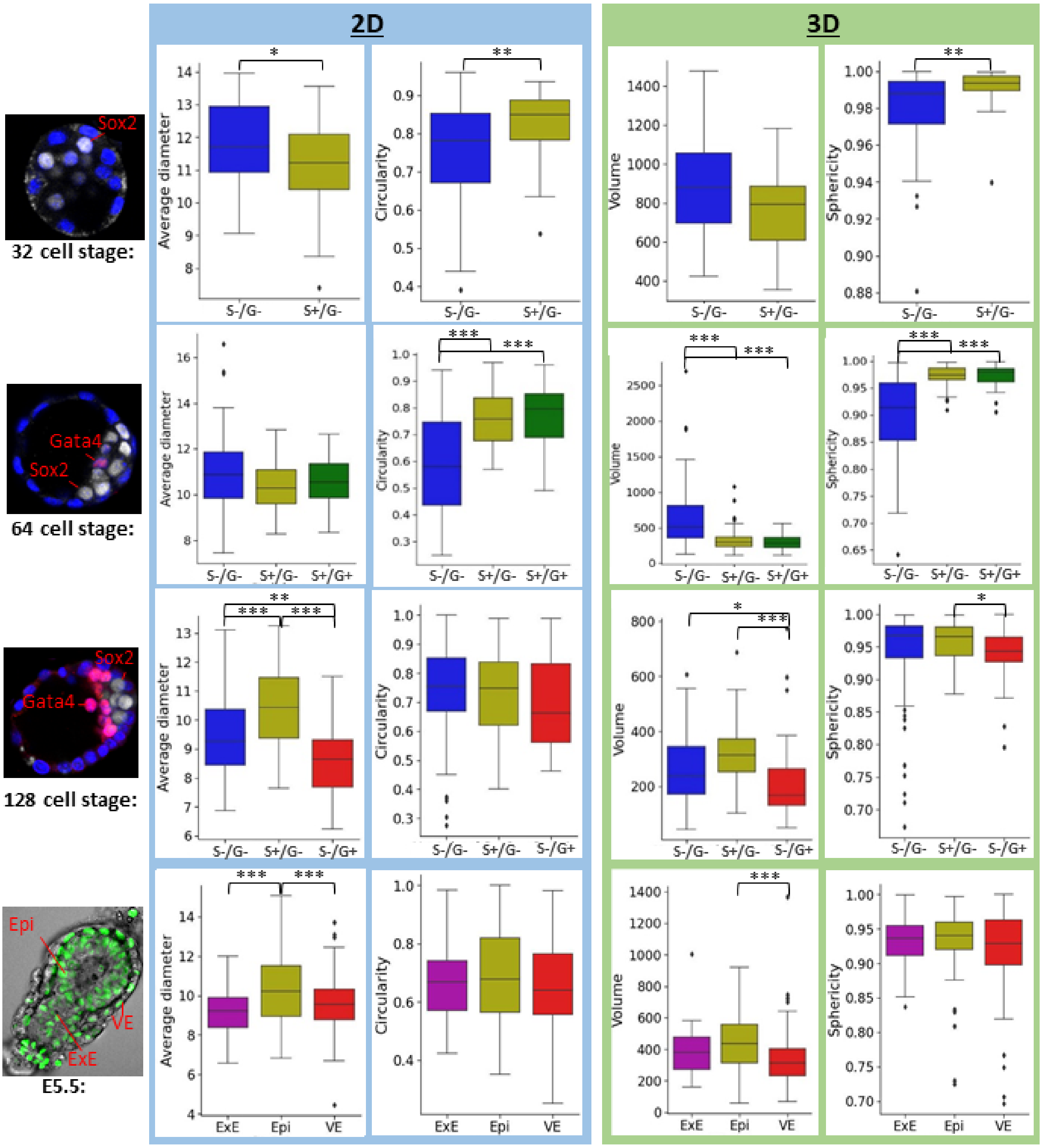


Figure 2: Variations in nuclei dimensions across cell lineage. Average diameter and circularity (calculated from 2D measurements) and volume and sphericity (calculated from 3D measurements) are shown for each lineage in each stage from 32 cell to E5.5. Embryos up to 128 cell are fixed and stained so lineage could be inferred; SOX2 negative GATA4 negative (S-/G-) nuclei are likely TE, SOX2 positive GATA4 negative (S+/G-) marks ICM/EPI nuclei, SOX2 positive GATA4 positive (S+/G+) marks PrE precursors, and SOX2 negative, Gata4 positive (S-/G+) marks PrE nuclei. At E5.5 live embryos were used as lineage could be determined from position in embryo. All box plots represent all individual nuclei from the specified lineage in all embryos of that developmental stage. 17 embryos were used in total. Overall p values for average diameter were 0.114 (from Kruksall-Wallis), $1.14\text{x}10^{-6}$ (from Kruksall-Wallis), and $4.44\text{x}10^{-5}$ (from ANOVA) for 64 cell, 128 cell and E5.5 respectively. Overall p values for circularity were $1.39\text{x}10^{-9}$ (from Kruskall-Wallis), 0.175 (from Kruksall-Wallis), and 0.234 (from Kruksall-Wallis) for 64 cell, 128 cell and E5.5 respectively. Overall p values for volume were $3.40\text{x}10^{-10}$ (from Kruksall-Wallis), 0.0245 (from Kruksall-Wallis), and $7.32\text{x}10^{-4}$ (from Kruksall-Wallis), for 64 cell, 128 cell and E5.5 respectively. Overall p values for sphericity were $2.26\text{x}10^{-15}$ (from Kruksall-Wallis), $5.98\text{x}10^{-5}$ (from Kruksall-Wallis), and 0.448 (from Kruksall-Wallis) for 64 cell, 128 cell and E5.5 respectively.

At the 64 cell stage the GATA4, SOX2 negative TE cells are significantly less circular and spherical than the SOX2 positive nuclei. Within the SOX2 positive subset there is no difference in the circularity or sphericity between the GATA4 positive and GATA4 negative nuclei. Nuclear volume was also significantly higher in TE nuclei than SOX2 positive nuclei, but again there was no difference between GATA4 positive and negative nuclei amongst SOX2 positive nuclei. The trend seen in average diameter is the same as that seen in volume, however for average diameter there are no significance differences. Our results suggest that the TE nuclei are still more elliptical than the ICM nuclei at this stage, and possibly still larger.

At the 128 cell stage the SOX2 positive nuclei are significantly larger than the GATA4 positive nuclei and the GATA4, SOX2 negative TE nuclei in 2D. The 3D measurements show the same trend, however here SOX2 positive nuclei are only significantly bigger than the GATA4 positive nuclei, and not the TE nuclei. The GATA4 positive nuclei are also significantly smaller than the TE nuclei in both 2D and 3D. The nuclei less vary less in ellipticity at this stage, the only significant difference found is that the GATA4 positive nuclei are less spherical than the SOX2 positive nuclei. The 2D measurements also show that GATA4 positive nuclei are less circular than SOX2 positive nuclei, although not significantly. Overall, these results suggest that the EPI nuclei are the largest and the PrE nuclei may be the smallest, with the PrE nuclei possibly being slightly less spherical than the EPI nuclei.

At E5.5, our results suggest that EPI nuclei are once again the largest, with the difference being significant when comparing to VE nuclei in both 2D and 3D and when comparing to ExE nuclei in 2D. EPI nuclei are also larger than ExE nuclei in 3D, however there is no significant difference. There are no noticeable differences between the circularity/sphericity of nuclei from different lineages.

Overall, we have found that at 32-64 cell stage there are differences in the circularity and size of the TE and ICM nuclei, but not between EPI and PrE precursors. TE cells are more elliptical than ICM cells and probably also bigger. At 128 cell stage and E5.5 the strongest differences between lineage is in nuclei size, with the EPI nuclei appearing to be the largest.

Early to mid-blastocyst TE nuclei have oblate shape

Following our finding that the TE nuclei are less circular/spherical than the ICM nuclei at 32-64 cell stage, we were interested to investigate this difference further. Spheroids can have two main types of shape; oblate (pancake shaped) or prolate (rugby ball shaped). Therefore, we re-used our 3D measurements of the 32 and 64 cell embryos (the stages that had shown the strongest differences in sphericity between lineages) to calculate oblateness and prolateness for each nuclei for these stages. The results, shown in Fig. 3a-b, show that the ICM nuclei are significantly less oblate than the TE nuclei for both stages, and there is no difference between GATA4 positive and negative nuclei in the ICM. This is similar to the trend seen in sphericity (although reversed because for sphericity a lower number indicates greater deviation from a sphere). Whereas for prolateness, there is no difference between TE and ICM nuclei at 32 cell stage and the TE is actually less prolate at 64 cell stage. These results suggest that the difference observed between TE and ICM nuclei is due to the TE cells becoming oblate rather than prolate.

We hypothesised that this could be due to pressure from the expanding cavity causing the cells to flatten. To test this theory, we looked at the correlation between Z slice and circularity for 2D measurements of TE nuclei for three 64 cell stage embryos, the results are shown in Fig 3c. The results show that the TE cells appear more elliptical around the central planes rather than at the top or bottom planes. This is what we would expect to see if pressure from the cavity is causing oblateness as in the central planes we are viewing the TE cells perpendicular to the direction of the pressure that would be exerted on them. Therefore, these results support our theory that pressure from the cavity is driving oblateness in TE nuclei.

Unsupervised clustering of nuclei shapes

All shape measurements so far have worked under the assumption that nuclei are an elliptical (in 2D) or spheroid (in 3D) shape, however it is possible that there may be a broader range of cell nuclei shapes that cannot be described by just maximum and minimum diameters. Therefore, we decided to further investigate the distribution of nuclei shapes using unsupervised clustering. We chose to use the 128 cell stage for this analysis, as we had noticed a potential variety of irregular shapes at this stage.

We prepared a dataset of images of 2D nuclei outlines by tracing around the outside of the nucleus in Imagej using the nuclear Hoechst staining and then cropping around the outline, see Fig. S2 for examples. The inside area was cleared and the outside area filled to give a binary image, this was to encourage the clustering algorithm to focus on the shape rather than staining intensity/variation. K-means clustering was performed on high level convolutional neural network (CNN) features for each image. We first pre-trained the CNN model to predict circularity and then used the features in the 2nd to last layer of this model as inputs to the K-means algorithm (full details in methods). The pre-training step was used as an attempt to create high level features that contained information on the shape of the nuclei rather than other factors such as orientation. We repeated K-means clustering with *k* values from 2 to 10.

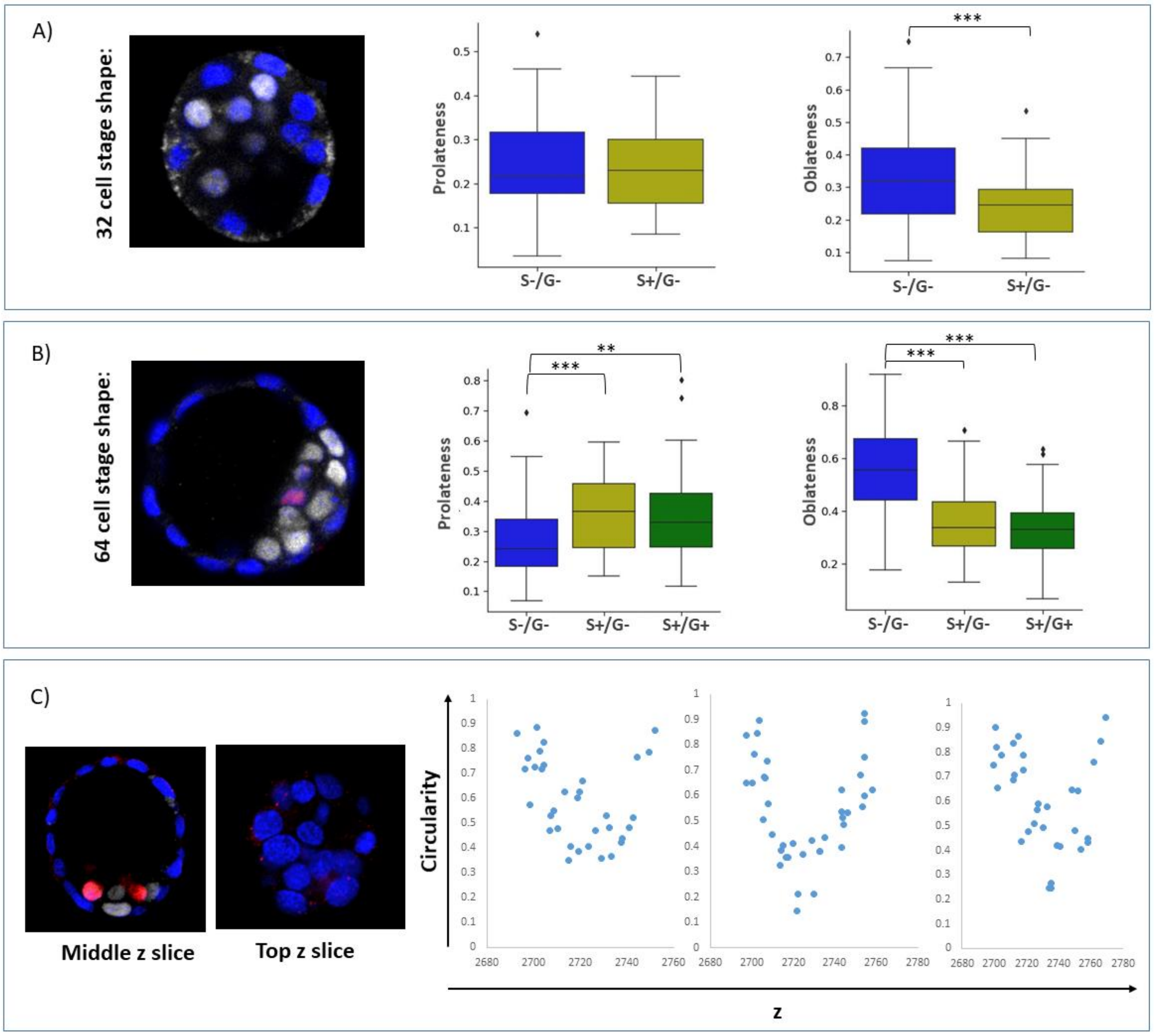


Figure 3: Investigation into spheroid type A) The prolateness and oblateness of nuclei of the 32 cell stage embryo. Measurements were taken from four embryos in total. B) The prolateness and oblateness of nuclei of the 64 cell stage embryo. An overall p value of 9.73x10^{-5} (from Kruksall-Wallis) was found for prolateness and 1.13x10^{-13} (from Kruksall-Wallis) for oblateness. Measurements were taken from four embryos in total. C) circularity of nuclei vs z slice for all TE (SOX2 negative GATA4 negative) nuclei in three 64 cell stage embryos, with example illustration of top and middle slice. All results are for fixed and immunostained embryos. S-/G- refers to SOX2 negative GATA4 negative, S+/G- refers to SOX2 positive GATA4 negative, and S+/G+ refers to SOX2 positive GAT4 positive. All box plots represent all individual nuclei from the specified lineage in all embryos of that developmental stage.

First, we created plots of distortion (the average squared distance of a sample to the cluster centre) and inertia (the sum of squared distances of samples to their closest cluster center) against *k* (Fig 4a). This allowed us to use the elbow method, which states that the point after which distortion/inertia start

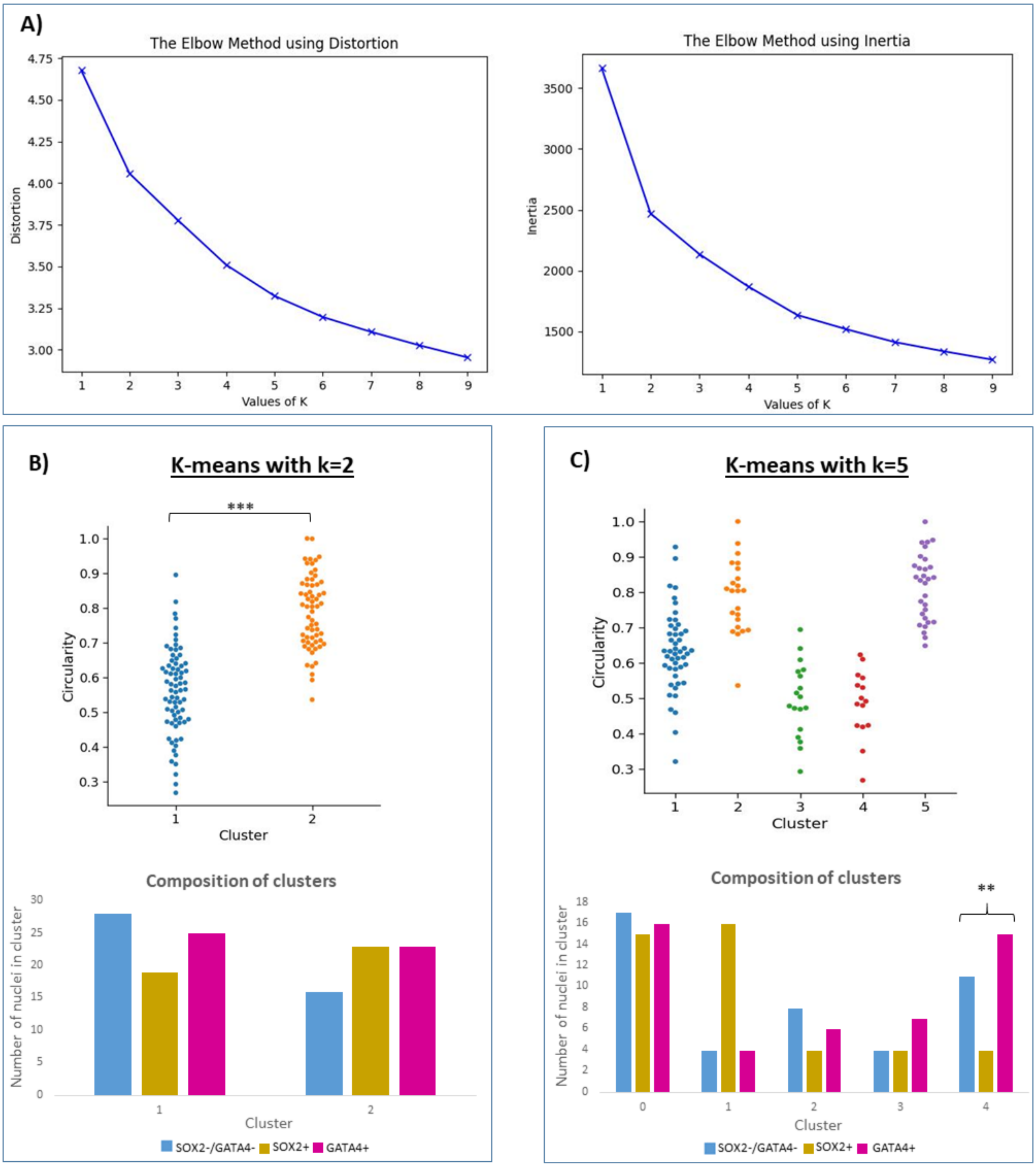


Figure 4: Unsupervised clustering of 128 cell stage nuclei. The results of K-means clustering on images of cropped outlines of nuclei in 128 cell embryos. A) Plots of distortion versus *k* number and inertia vs *k* number produced to determine optimal *k* value using the elbow method. B) The distribution of circularity values in each cluster and lineage

composition of each cluster when *k*=2. The difference between circularity in each cluster had a p value of $5.55\text{x}10^{-20}$ (from students' t-test) . The overall p value for lineage composition of the clusters was 0.222 (from Chi squared) C) The distribution of circularity values in each cluster and lineage composition of each cluster when *k*=5. The difference between circularity in each cluster had a p value of $4.21\text{x}10^{-17}$(from ANOVA). The overall p value for lineage composition of the clusters was $8.26\text{x}10^{-4}$ (from Chi-squared).

decreasing with *k* in a linear fashion is the optimal cluster number. Although there is not an obvious inflection point on these plots, it does appear that both *k*=2 and *k*=5 may be good choices. Although *k*=2 is perhaps a more marked inflection point, this is likely to just reflect the two 'circular' and 'elliptical' groups in the pre-training step, and the purpose of this exercise is to investigate variation in shape beyond circularity. Therefore, we chose to investigate the clusters produced with *k*=2 to check the clustering is working as we would expect and also *k*=5 to see the potential variation in shapes in more detail.

The circularity of each nucleus for *k*=2 is shown in Fig. 4b. As expected, one group is significantly more elliptical than the other. Also shown is the composition of each group in terms of the lineages, we found no significant difference between the two clusters. This is what we would expect as our earlier results did not find a difference between circularity between the three different lineages at this stage. Together, these results suggest that the clustering algorithm is using shape to assign nuclei to clusters, as was desired.

Next, we carried out the same analysis for *k*=5, as shown in Fig. 4c. The clusters appeared to vary in circularity, and an ANOVA test confirmed that there is significant variation between the clusters. This suggests that the algorithm is still using shape to perform the clustering. There does also appear to be some variation in the proportion of each lineage in the clusters. In particular, cluster 2 seems to be mainly composed of SOX2 positive nuclei while cluster 5 seems to have very few SOX2 positive nuclei in comparison to GATA4 positive and TE (double negative) nuclei, although only cluster 5 was significantly different from the expected distribution. Interestingly, these two clusters have very similar distributions of circularity values, with both being more circular than the other three clusters. This suggests that these results are not due to EPI nuclei simply being more or less circular than other nuclei, but may instead indicate another difference in shape. However, more data would be needed to draw any firm conclusions.

<u>Discussion</u>

In this study we have conducted a comprehensive analysis into the variation in nuclei size and shape in the early mouse embryo from the 8-cell stage to peri-implantation. Very little was previously known about the relation of nuclei dimensions to embryonic stage and lineage in mammalian embryos. We have found new correlations between nuclei size/shape and cell lineage at all stages from 32 cell to E5.5 and have also discovered the point in development at which nuclei size begins to increase following a decrease in line with early cell cleavages.

We have used two approaches for nuclei measurements; 2D and 3D, both of which have associated strengths and weaknesses. The 2D measurements have less measurement error due to higher resolution in the XY plane, and they are also less prone to subjectivity as the edge of the nuclei appears sharper in the central and most in focus Z slice. However, 2D measurements have the obvious drawback of approximating a 3D object in two dimensions. Meanwhile, 3D measurements have more associated error due to lower Z resolution and are prone to subjectivity as it can sometimes be difficult to ascertain the correct edge of the nuclei, especially in crowded locations such as the ICM where nuclei appear to overlap. Although both approaches have limitations, the fact that the same trends are seen in both methods strengthens our findings.

The analysis of nuclei size and shape across development suggested that nuclei diameter decreases from 8 cell to 128 cell, then increases from 128 cell stage to E5.5, and circularity decreases from 16 cell stage. The same trends in circularity and diameter were seen for disaggregated and intact embryos, however the trend was stronger in intact embryos. This may be due to sample size, more cells were included for most stages in the intact embryos, especially the E4.5 stage which differed from the general trend for circularity.

The pattern of nuclei diameter over development is the same as that found for cell size vs developmental stage by Forsyth et al. (2021)[17]. Nuclei size vs developmental stage has also been studied by Tsichlaki and Fitzharris (2016)[16], who found that the nuclei decreased in size from zygote to blastocyst, agreeing with our findings. As far as we are aware we are the first to show that nuclei size begins to increase at E5.5. The change in shape of nuclei over development has not previously been reported, however blastomere sphericity has been shown to decrease from the 2 to 64 cell stage by Royer et al. (2020)[18]. They hypothesised that this was due to an increase in the variety of cell shapes as the embryo develops[18].

Our analysis into how nuclei size and shape varied across lineage found that the most strongly marked difference at 128 cell stage and E5.5 was in nuclei size with the EPI nuclei significantly bigger than the extra-embryonic lineages. Our results also showed that the PrE nuclei were the smallest at 128 cell stage, and suggested that PrE nuclei are possibly less round that EPI nuclei, although this was only statistically significant in 3D. At the 32 and 64 cell stage the TE nuclei were found to be significantly less circular/spherical and larger than the ICM nuclei, however the latter was only significant in the 2D measurements at 32 cell stage and 3D measurements at the 64 cell stage. We found that the lower sphericity of the TE nuclei at 32 cell and 64 cell stage was due to the TE nuclei becoming oblate, and provided evidence that this was due to pressure from the cavity. Our clustering analysis suggested that there may be specific types of shapes that are more or less common in different lineages, however more data would be needed to confirm this.

As far as we are aware, we are the first to report differences between the size and shape of nuclei across cell lineages in the early embryo. The only other study reporting differences relating to nuclei dimensions in the early embryo is Aiken et al (2004)[19], who found that nucleo-cytoplasmic ratio was significantly higher in the TE than the ICM at E3.5 and outer cells were larger than inner cells. They do not report nuclei volumes directly, however these results indicate agreement with our finding that TE nuclei are larger than ICM nuclei at E3.5.

If a standard distribution of sizes and shapes could be established it potentially could be possible to detect embryos that have developed abnormally just by looking at the nuclei shape and size. In particular, if links between particular clusters and lineage were to be confirmed this could be used to assess whether there might be a deficiency or surplus of cells in a lineage without the need for immunostaining. This information could potentially be useful for embryo assessment in IVF procedures where embryos are selected on a morphological basis.

Conclusion

In this study we have discovered trends in mouse nuclei sizes and shapes over early development, with roundness generally decreasing over time and size initially decreasing before starting to increase from the E4.5 to E5.5 stage. We have also found links between cell lineage and nuclei size and shape, in particular the TE nuclei appear to be larger and more elliptical than ICM nuclei at 32-64 cell stage and the EPI nuclei have been found to be larger than nuclei of extra-embryonic lineages at 128 cell stage and E5.5.

Methods

Collection of mouse embryos

Mice were housed in the Biological Service Facility (BSF), University of Manchester, under a 12-hour light cycle. Two mouse strains were used; CD-1 outbred (Jackson Laboratories) and CAG::H2B-EGFP transgenic mice, the latter allow visualisation of chromatin through the fluorescent histone protein[20]. Mating was detected by the presence of a vaginal plug, 12pm on the day of plug observance was estimated to be embryonic day 0.5 (E0.5). Mice were sacrificed on days E2.5-E5.5 to obtain embryos at a variety of developmental stages, and embryos were collected by flushing the oviducts (E2.5) or the uterus horns (E3.5 onwards) using warm home-made M2 medium[21].

The mice were bred on project license P08B76E2B, protocol 4 and the license 70/08858, protocol 4, and all husbandry and handling methods conformed to the regulations set out by the UK Home Office's Animals (Scientific Procedures) Act 1986. The mice were humanely euthanised in accordance with Schedule 1 of the UK Animals (Scientific Procedures) Act 1986. Ethical approval for the euthanasia of animals used in this study was granted to the project submitted by Berenika Plusa by the University of Manchester Animal Welfare and Ethical Review Body on the 10/03/2017.

Zona pellucida removal

For all the immunostained and disaggregated embryos the zona pellucida was removed by moving embryos between several drops of warmed acid Tyrode's solution (Sigma Aldrich). This took place under careful visual inspection; embryos were quickly removed as soon as the zona pellucida had fully dissolved and were then washed and left to recover in M2 for a minimum of 20 minutes.

Disaggregation

Two protocols were used to disaggregate the embryos; a calcium free treatment for developmental stages up to and including E4.5 and trypsin treatment for E4.5 and E5.5 embryos. Both protocols were carried out by Forsyth et al. (2021)[17]. Groups of E4.5 embryos were disaggregated using both methods to ensure the disaggregation method did not affect the results, no significant differences in cell diameters were observed across the 2 treatments[17].

Fixation and Immunostaining

Embryos were fixed and stained so that lineages could be distinguished. The Epi was marked by SOX2 as it is a well known marker of pluropoteniency in embryos[22,23]. The PrE was marked by the first exclusive PrE marker; Gata4[24]. No marker was used for TE as it could be distinguished by position and lack of other markers.

All embryos intended to be immunostained were first fixed in 4% para-formaldehyde (Sigma Aldrich) in PBS with 0.1% Tween-20 (Sigma) and 0.01% Triton X-100 (Fluka) for 20 minutes. Following this, the embryos were washed in PBS and stored (also in PBS) at 4°C until removal for immunostaining. The immunostaining was carried out following the methodology described in Plusa et al. (2008)[24]. Permabilisation was performed by leaving embryos in 0.55% Triton-X 100 in PBS for 20 minutes. Before exposure to both primary and secondary antibodies the embryos were blocked in 10% donkey serum (Sigma Aldrich) in PBS for 40 minutes. The primary antibodies used were anti-Gata4 (Santa Cruz Biotechnology, 1:200) and anti-Sox2 (ThermoFischer Scientific, 1:100) overnight at 4°C. Secondary Alexa Fluor (Invitrogen) conjugated antibodies were used (1:500) for 1 hour at 4°C. Finally, embryos were incubated in Hoechst 33342 (Sigma Aldrich) at a concentration of 1:1000 in PBX (PBS +0.1% Triton-X 100) for at least 30 minutes at 4°C in order to visualise the nuclei.

Image acquisition

All imaging was performed by the Nikon A1 inverted confocal microscope. Embryos were placed on glass-bottom dishes (MatTek) under mineral oil. Sections were imaged at micrometer intervals with Diode 405 nm, Argon 488 nm, HeNe 546 nm, and HeNe 647 nm lasers used to excite flurorophores.

Image analysis

All image analysis was carried out using ImageJ and IMARIS (Bitplane). For embryos up to stage E4.5 a cell count was calculated using the spot detector in IMARIS to detect nuclei and accuracy of spot detection was inspected manually. Measurements were performed using the nuclear Hoescht 33342 staining for fixed embryos, H2B-GFP fluorescence for live and disaggregated embryos. All 2D and manual 3D measurements were taken using the ruler tool in IMARIS, with 2D measurements always taken in the z slice with maximum cross section. Automated 3D measurements were taken using the surfaces tool in IMARIS.

For fixed embryos, presence of GATA4 or SOX2 staining was recorded for each nuclei. For live E5.5 embryos the brightfield images were used to identify cell lineage.

The preparation of images for clustering was carried out using ImageJ. The ellipticity of nuclei was measured using the ruler tool and the outline of each nuclei was then traced using the polygon tool. Images of individual nuclei were then created by cropping around the polygon and filling the inside of the polygon and clearing the outside using ImageJ. All images were resized to 224x224 using the python imaging library (PIL).

3D metric calculations

For each nuclei measured in 3D we obtained 3 perpendicular measurements, which we refer to as a, b, and c, where a≤b≤c. The volume was calculated using the formula for an ellipsoid:

$$V = 4\pi \mathrm{abc}/3$$

We calculated sphericity from the formula:

$$Sphericity = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}$$

Where V is the volume as calculated above and A is the surface area, calculated by the approximate surface area of an ellipsoid:

$$A = 4\pi(\frac{(ab)^{1.6} + (ac)^{1.6} + (bc)^{1.6}}{3})^{\frac{1}{1.6}}$$

Oblateness was calculated as:

$$Oblateness = \frac{2a^2}{a^2 + b^2} \cdot (1 - \frac{a^2 + b^2}{2c^2})$$

And prolateness was calculated as:

$$prolateness = \frac{2b^2}{b^2 + c^2} \cdot (1 - \frac{2a^2}{b^2 + c^2})$$

Unsupervised clustering algorithm

k-means clustering was performed on high-level CNN features. To obtain features that would be appropriate for clustering based on shape a supervised model was first trained to predict nuclei circularity. For this, we used the MobilenetV2 model[25] with weights pre-trained on the ImageNet database[26]. An extra hidden layer was added before the final layer and all previous layers were fixed. We divided our dataset into two classes; 'elliptical' and circular' based on the nuclei circularity, using a circularity of 0.66 as the threshold as this was the median value. We then trained the model to predict these classes using all images as training data, and used a base learning rate (BLR) of 0.0001, a drop out of 0.5 and the cross entropy loss function. Once the model had been trained we saved the values of the hidden layer features for each image, and used these as the features for the clustering algorithm.

For the unsupervised clustering, we used the scikit-learn Kmeans function in python. The *k* value was varied from 2 to 10, while all other variables were set to default.

Statistics

Continuous data was tested for normality using the D'Agostino & Pearson test. For cases with two datasets being compared a students t-test was used to test for significance if the data was normally distributed and a Mann Whitney test was performed if the data was not normally distributed. If more than two datasets were being compared and they were normally distributed a one-way ANOVA test was used followed by a post-hoc test (HSD) if the result was significant, if the data was not normally distributed a Kruksall-Wallis test was used, followed by a post-hoc test (Dunns) if the result was significant. For the categorical data, significance was tested using chi-squared tests.

For all significance tests, results were determined to be significanct if $p<0.05$. For all results shown graphically, * corresponds to $p<0.05$, ** corresponds to $p<0.01$, and *** corresponds to $p<0.001$

References

1 Aplin, J. D. & Ruane, P. T. Embryo-epithelium interactions during implantation at a glance. *Journal of cell science* **130**, 15-22 (2017). https://doi.org:10.1242/jcs.175943

2 Hemberger, M., Hanna, C. W. & Dean, W. Mechanisms of early placental development in mouse and humans. *Nature reviews. Genetics* **21**, 27-43 (2020). https://doi.org:10.1038/s41576-019-0169-4

3 Johnson, M. H. & Ziomek, C. A. The foundation of two distinct cell lineages within the mouse morula. *Cell* **24**, 71-80 (1981). https://doi.org:10.1016/0092-8674(81)90502-X

4 Dickson, A. D. The form of the mouse blastocyst. *Journal of anatomy* **100**, 335-348 (1966).

5 Robertson, E. J. & Arnold, S. J. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nature reviews. Molecular cell biology* **10**, 91-103 (2009). https://doi.org:10.1038/nrm2618

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

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

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

9 Nasiri, N. & Eftekhari-Yazdi, P. An overview of the available methods for morphological scoring of pre-Implantation embryos in in vitro fertilization. *Cell journal (Yakhteh)* **16**, 392-405 (2015).

10 Barberet, J. *et al.* Can novel early non-invasive biomarkers of embryo quality be identified with time-lapse imaging to predict live birth? *Human reproduction (Oxford)* **34**, 1439-1449 (2019). https://doi.org:10.1093/humrep/dez085

11 Otsuki, J. *et al.* Noninvasive embryo selection: kinetic analysis of female and male pronuclear development to predict embryo quality and potential to produce live birth. *Fertility and sterility* **112**, 874-881 (2019). https://doi.org:10.1016/j.fertnstert.2019.07.015

12 Manor, D., Drugan, A., Stein, D., Pillar, M. & Itskovitz-Eldor, J. Unequal pronuclear size : A powerful predictor of embryonic chromosome anomalies. *Journal of assisted reproduction and genetics* **16**, 385-389 (1999). https://doi.org:10.1023/A:1020550115345

13 Nagy, Z. P. *et al.* Pronuclear morphology evaluation with subsequent evaluation of embryo morphology significantly increases implantation rates. *Fertility and sterility* **80**, 67-74 (2003). https://doi.org:10.1016/S0015-0282(03)00569-7

14 Sadowy, S., Tomkin, G., Munné, S., Ferrara-Congedo, T. & Cohen, J. Impaired development of zygotes with uneven pronuclear size. *Zygote (Cambridge)* **6**, 137-141 (1998). https://doi.org:10.1017/S0967199498000057

15 Webster, M., Witkin, K. L. & Cohen-Fix, O. Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *Journal of cell science* **122**, 1477-1486 (2009). https://doi.org:10.1242/jcs.037333

16 Tsichlaki, E. & Fitzharris, G. Nucleus downscaling in mouse embryos is regulated by cooperative developmental and geometric programs. *Scientific reports* **6**, 28040-28040 (2016). https://doi.org:10.1038/srep28040

17 Forsyth, J. E. *et al.* IVEN: A quantitative tool to describe 3D cell position and neighbourhood reveals architectural changes in FGF4-treated preimplantation embryos. *PLoS biology* **19**, e3001345-e3001345 (2021). https://doi.org:10.1371/journal.pbio.3001345

18 Royer, C. *et al.* Establishment of a relationship between blastomere geometry and YAP localisation during compaction. *Development (Cambridge)* **147** (2020). https://doi.org:10.1242/dev.189449

19 Aiken, C. E. M., Swoboda, P. P. L., Skepper, J. N. & Johnson, M. H. The direct measurement of embryogenic volume and nucleo-cytoplasmic ratio during mouse pre-implantation development. *Reproduction (Cambridge, England)* **128**, 527-535 (2004). https://doi.org:10.1530/rep.1.00281

20 Hadjantonakis, A.-K. & Papaioannou, V. E. Dynamic in vivo imaging and cell tracking using a histone fluorescent protein fusion in mice. *BMC biotechnology* **4**, 33-33 (2004). https://doi.org:10.1186/1472-6750-4-33

21 Grabarek, J. B. & Plusa, B. Live imaging of primitive endoderm precursors in the mouse blastocyst. *Methods Mol Biol* **916**, 275-285 (2012). https://doi.org:10.1007/978-1-61779-980-8_21

22 Avilion, A. A. *et al.* Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & development* **17**, 126-140 (2003). https://doi.org:10.1101/gad.224503

23 Wicklow, E. *et al.* HIPPO Pathway Members Restrict SOX2 to the Inner Cell Mass Where It Promotes ICM Fates in the Mouse Blastocyst: e1004618. *PLoS genetics* **10** (2014). https://doi.org:10.1371/journal.pgen.1004618

24 Plusa, B., Piliszek, A., Frankenberg, S., Artus, J. & Hadjantonakis, A.-K. Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development (Cambridge)* **135**, 3081-3091 (2008). https://doi.org:10.1242/dev.021519

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

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

Supplementary Materials

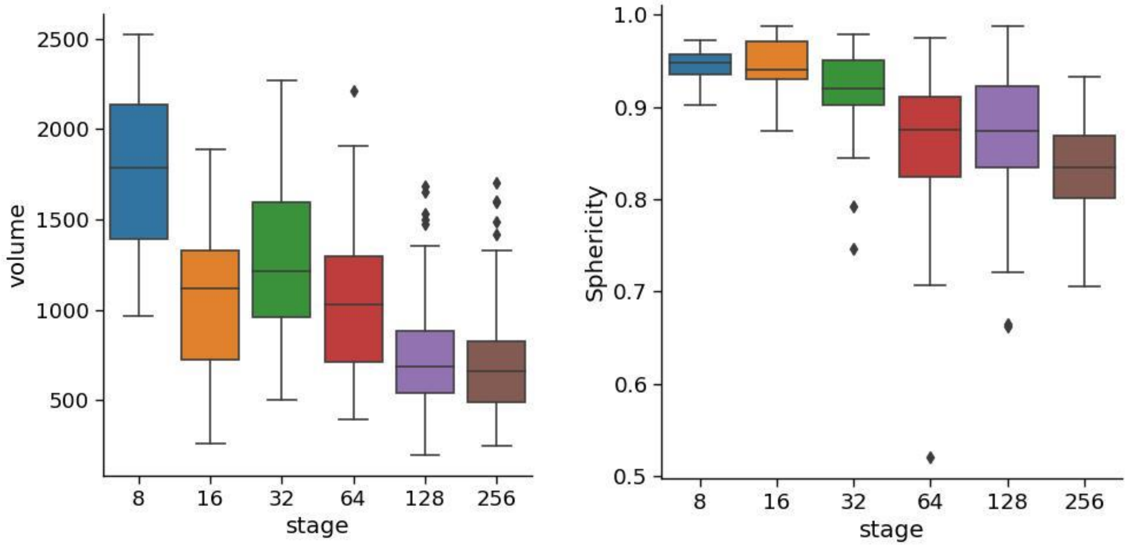


Figure S1: Volume and Sphericity across developmental stage measured using IMARIS surface tool

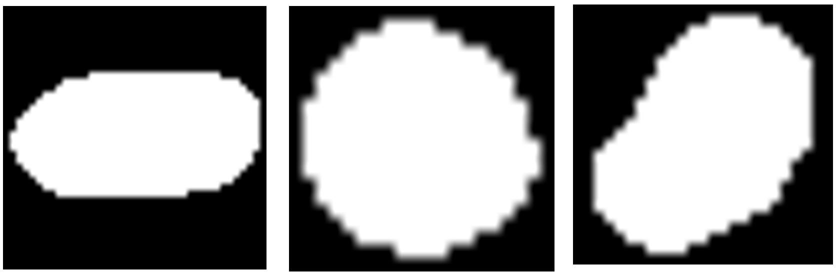


Figure S2: Examples of cropped nuclei images used as inputs to K-means algorithm. Nuclei outlines were traced using the polygon tool on ImageJ and then converted to binary images.

Chapter 5:

General Discussion

Despite much progress in IVF treatment over the last few decades, challenges remain in improving success rates and lowering risks to the mother and baby. Currently, success rates are still far from satisfactory, ranging from 32% for patients under 35 to under 5% for patients over 43[2]. One of the main challenges that remains to be addressed is selecting the best embryo to be transferred, as it is still not known exactly what healthy pre-implantation development should look like[6,61]. Multiple publications have explored potential markers of embryo viability including PN appearance[61-65], cell divisions[9,72-75] and cytoplasm morphology[60,66,67]. For some of the investigated features, contradictory conclusions have been reached, whereas the use of others has been limited. Importantly, the link between embryo viability and sub-cellular features, such as size and shape of nuclei, has not been explored beyond the first two embryonic cycles. Machine Learning (ML) approaches for medical purposes are undergoing rapid progress and potentially offer promising solutions to some of the problems related to IVF embryo selection process, especially if combined with human expert knowledge[4,132]. Ideally, ML models should be tailor-able to individual clinics and applicable to a range of transfer strategies, as earlier transfers may be beneficial to avoid the potential risks associated with blastocyst culture[41-44,140].

The overarching aim of this thesis was to contribute to safer and more successful IVF procedures. In particular, the research conducted during this PhD has tackled several current challenges in IVF including subjectivity in embryo selection, the lack of ML algorithms available for early stage embryo assessment, and the inattention to potentially important sub-cellular features when assessing embryo quality. Together, the three papers presented in this thesis work towards solutions to these amongst other issues. The paper presented in chapter 2 developed CNN models that were capable of automatically assessing the chance of live birth based on various stages of embryonic development. The paper in chapter 3 then combined these models with manually assessed features to further improve live birth prediction at early embryonic stages. Finally, the paper in chapter 4 aimed to deepen our understanding of the pre-implantation process by investigating normal development of the nucleus in mouse embryos, it is hoped that this could pave the way for inclusion of sub-cellular features like nuclear shape and size (past zygote stage) in embryo assessment procedures.

The paper in chapter 2, *'Deep learning pipeline reveals key moments in human embryonic development predictive of live birth in IVF'*, addressed several weaknesses in current approaches to DL-based assessment of embryo viability. Firstly, live birth was predicted directly rather using more commonly used intermediate outcomes. Secondly, we presented the first CNN models to predict live birth based solely on individual pre-blastocyst stages, allowing for earlier embryo selection. Thirdly, we showed it was possible to train a CNN to predict live birth to the same standard as experienced embryologists using data from a single clinic.

The study in chapter 2 took a different approach to the usual embryo assessment procedure of choosing frames by video time-point. Instead, we picked specific well-defined moments in development as reference points and used these to find the optimal developmental moments for predicting live birth. One advantage of our approach is that it ensured we were always comparing embryos that were actually at the same developmental stage, reducing unnecessary variation that would act as noise. Another advantage is that it allowed the pre-implantation time period to be thoroughly explored to reveal the best moments for live birth prediction, whereas choosing frames based on arbitrary time points such as 'day 3' may miss crucial moments of development where viability is most apparent. Our work is the first to identify specific moments that are optimal for live birth prediction, these findings could be useful both for future ML studies and more general embryologist based embryo assessment.

The performance of the blastocyst model was found to be similar to or better than models developed in the few previous studies that had also predicted live birth[136-138]. Our work was the first to employ DL to make live birth predictions solely from pre-blastocyst stages, so no direct comparisons to earlier studies are possible, however other researchers have explored the use of early stages for other purposes. Lee et. al predicted ploidy by taking images from various stretches of development, including just day 1 for which they had an AUC of 0.58[131]. We achieved a slightly higher AUC for our PN model despite having the more difficult classification problem of transfer outcome. This suggests that our strategies of extra transfer learning and carefully chosen frame extraction time-point have improved model performance and are important approaches to consider when training an ML model to assess embryo viability at early embryonic stages.

In the final section of the paper in chapter 2 we provided evidence that the pre-blastocyst models may add important information to selection at blastocyst stage, allowing for the best blastocyst to be selected even amongst those that may appear the same quality. Future avenues of research would be further combinations of the CNN models, and possibly incorporation of other manually assessed features. In order to identify manually assessed features that could strengthen the performance of IVF embryo assessment in conjunction with our CNN models, we performed work that was included in the paper presented in chapter 3.

The paper in chapter 3, '*Combined deep learning and human annotation approach allows cleavage stage assessment of human embryos',* first identified morphological features that were significantly different between successful (live birth) and unsuccessful embryos (transferred but no pregnancy) and then combined these with predictions from the CNN models. Here we focused on early stages (using only information up to cleavage stage) and got improved performance from this combination approach than what was achieved from any pre-blastocyst models individually, with a test set ROC AUC of up to 0.64. This strengthens our findings from the previous paper that early stages can be used for embryo selection, and provides the possibility of a more diverse treatment regimes that can allow for earlier transfers and/or cryopreservation, minimising the possible side effects of prolonged blastocyst culture.

The investigation into potential morphological quality markers led to us being the first (to our knowledge) to show that the presence of a 'rough patch' at zygote stage and degree of morula compaction are both significantly correlated to live birth. We also investigated two features for which the link to embryo viability is currently subject to dispute; zygote ZP thickness variation[69-71] and vacuoles[61,67], finding that only the latter was correlated with live birth. While we cannot know for certain why our results on vacuoles disagreed with the previous study by Barberet et al.[61], possible reasons are that they only included high-grade embryos that had been transferred and that they had a relatively small sample size.

In the paper in chapter 3 we also used unsupervised learning and dimensionality reduction techniques to better understand the dataset. Interestingly, we observed that while good quality embryos formed one bigger cluster, lower quality embryos were distributed between multiple, smaller clusters. Such a distribution seems to suggests that there may be a broad spectrum of developmental issues that can lead to embryo failure, while heathy embryos follow more similar developmental paths. This provides a greater insight into the challenge of embryo selection, as it suggests that there may be a range of developmental abnormalities to look out for when assessing embryo viability. Our findings here highlight the advantage of the approach taken in this thesis of examining various morphological features throughout development.

One of the major limitations of the work described in chapter 2 and chapter 3 papers is that the time-lapse videos used had only one z-slice and were of varying quality with not all features possible to assess in all videos. These limitations mean that sub-cellular features were hard to examine. Most studies investigating potential morphological viability markers are based on similar datasets and as a result there is very little knowledge on the correlation of sub-cellular features and embryo viability. To overcome this multiple z-slices for each time-point could be exported from the embryoscope and a CNN adapted to 3D images, such as a model with RNN-CNN architecture could be trained to predict viability from these more detailed images.

Another limitation is that all the embryos in the dataset used for chapters 2 and 3 were produced via ICSI. Therefore, it is possible that the models trained here may not perform as well on embryos created via traditional IVF. As many embryos in IVF clinics are created via traditional IVF and not all clinics offer ICSI it is important that the models developed here should next be tested on a dataset of embryos created by traditional IVF.

Another limitation is that the models developed help us to predict whether or not an embryo is likely to result in live birth, however they do not inform us of the reason for embryo failure. The reason for infertility is not always known[141], which makes it harder to advise the patients on the best next steps to take or tailor treatment towards specific infertility issues. In order to better diagnose the reasons for embryo failure it is important to first understand normal pre-implantation development. This can be difficult when using human embryos from an IVF clinic as the majority of embryos come from patients struggling with infertility so they are more likely to have various developmental abnormalities. A potential strategy to overcome this would be to study embryos donated to research from patients known to not have fertility related issues (e.g patients undergoing IVF for embryo genetic screening to prevent genetic diseases) to get a clearer understanding of healthy embryo development.

Studying features important for assessment of viability in human embryos is often restricted due to unavailability of high-quality material and ethical constrains. At the same time, mouse embryos are a well established model to study pre-implantation mammalian development. Due to the ease of accessibility, relatively low costs, and high quality (unlike in humans, the vast majority of mouse embryos develop to term), mouse embryos are a particularly useful model to investigate what the healthy embryo looks like.

Therefore, in our final paper, '*Nuclei dimensions shown to correlate with developmental stage and cell lineage in the early mouse embryo'* we used mouse embryos to investigate how size and shape of nuclei change during preimplantation development. There is currently no knowledge of how nuclei size and shape (past zygote stage) might be related to embryo viability and the pattern of typical nuclei appearance across development and cell lineage has not yet been established. However, the size of the nucleus is known to be important to cellular function and development, with irregularities related to disease[142]. Therefore, nuclei size and shape may be a critical component currently missing in embryo assessment.

The images collected for this work allowed for a much more detailed examination of the nuclei than in IVF time-lapse videos as we were able to take many Z slices to get a 3D view of the embryo and we also included stained embryos for which the cell lineages could be inferred. From this dataset, we found that nuclei decrease in both size and sphericity throughout the pre-implantation period and TE nuclei are both larger and more elliptical than ICM nuclei at early to mid-blastocyst.

If similar trends can also be found in human embryos, then it is possible that any deviations could be used as indicators of poor embryo quality. For example, if nuclei are larger than expected at a particular stage this could mean that not enough divisions have occurred. The proportion of nuclei with certain dimensions could also potentially be useful for determining whether all lineages have formed correctly. Additionally, the clustering approach that we explored here could be particularly useful in describing the typical distribution of shapes that should be expected as it can incorporate the whole nuclei outline and does not require any prior shape approximations.

Another key finding from this paper was that 2D and 3D measurements produced the same general trends in size and shape. Therefore, we have provided evidence that 2D measurements are likely to be sufficient to describe the typical appearance of nuclei across developmental stage and cell lineage. This could be important information for any further investigation using human embryos, as obtaining 3D measurements from time-lapse videos recorded in IVF clinics may be challenging due to the typically small number of Z-slices taken.

Summary of findings

Chapter 2: *'Deep learning pipeline reveals key moments in human embryonic development predictive of live birth in IVF'*

-Optimal moments in development to predict live birth were revealed.

-CNN models were developed to predict live birth at various individual points of development from zygote to blastocyst.

-We demonstrated it was possible to train a CNN to predict live birth to a high performance using a single-clinic dataset.

-Early developmental stages were shown to add value to live birth prediction even when selecting embryos at the blastocyst stage.

Chapter 3: '*Combined deep learning and human annotation approach allows cleavage stage assessment of human embryos'*

-Novel morphological markers of viability were identified.

-CNN outputs and manually assessed features were combined by a variety of supervised learning algorithms to predict live birth at the cleavage stage with a ROC AUC of up to 0.64.

-Unsupervised clustering and dimensionality reduction revealed the presence of several clusters of lower quality embryos.

Chapter 4: '*Nuclei dimensions shown to correlate with developmental stage and cell lineage in early mouse embryos'*

-Mouse nuclei size was found to decrease until the E4.5 stage and then increase at E5.5, while the nuclei sphericity was found to generally decrease over development.

-Significant differences in nuclei size and shape across different lineages were found at all stages from 32 cell to E5.5, the most strongly marked being that the TE nuclei were more elliptical than the ICM nuclei at the 32-64 cell stage and the EPI nuclei were larger than the extra-embryonic nuclei at E4.5 and E5.5.

-Unsupervised clustering showed possible lineage-related clusters that differ in shape in a manner unrelated to circularity

Further work

The CNN models developed in chapter 2 were trained on a single-clinic dataset, with the intention being that they could be re-trained on datasets from other clinics to become tailored to their specific patient population. The next step would therefore be to check that the models obtain the same performances after re-training on datasets from other clinics. In particular, it is important to check that the moments of development found to have the greatest predictive power are equally useful on these different datasets. Similarly, the performance of the supervised ML algorithms developed in chapter 3 should be verified on other datasets.

In chapter 3 we discovered that combining manually assessed features with CNN outputs gave a better prediction of live birth than any of the pre-blastocyst CNN models alone. As the list of manual features included was by no means exhaustive, it is likely that this performance can be improved further with the inclusion of more features. This highlights two potential avenues of further work; identifying more features that can improve performance, and automating the detection of some or all of the current manual features. The latter is particularly important as most of the manually assessed features are prone to a high level of subjectivity, which could be solved by using advanced image analysis techniques and/or CNN models to detect whether each marker is present.

Once models have been further developed and verified, prospective trials would be necessary to assess whether they improve live birth success rates in practice. Additionally, it would be very interesting to investigate whether embryos in the same cohort tend to get similar model scores. This would inform us how capable the models are at ranking embryos in a cohort, which is the real challenge of embryo selection.

Another potential direction for further work would be to investigate whether the early stage models could be used to help determine whether good quality blastocysts fail due embryonic issues not visible at the blastocyst stage or due to maternal factors. This would require investigating whether any particular developmental abnormalities are associated with low early stage model scores and examining whether there are any correlations in these scores within cohorts.

Finally, in chapter 4 trends in the shape and size of nuclei in mouse embryos were discovered. The next step would be to check whether these trends also exist in human embryos. If the patterns are conserved in human embryos then possible links between deviation from normal nuclei appearance and transfer outcome could be investigated, the growing body of data from the recently available Embryoscope+ ™ makes this increasingly feasible as it produces high resolution time-lapse videos. In addition, further research is required to understand why the changes in nuclei size and shape take place. This could involve more experimentation using mouse embryos, such as interventions that perturb cell lineage formation to separate the effects of lineage and cell position on nuclei size and shape.

In conclusion, the research conducted as part of this thesis has contributed to our understanding of mammalian pre-implantation development and identified ML approaches capable of assisting with embryo selection. Our findings have raised many exciting further avenues of research that have the potential to led to advancements in IVF treatment.

Bibliography

1 Steptoe, P. C. & Edwards, R. G. BIRTH AFTER THE REIMPLANTATION OF A HUMAN EMBRYO. *The Lancet (British edition)* **312**, 366-366 (1978). https://doi.org:10.1016/S0140-6736(78)92957-4

2 Authority, H. F. a. E. (2021).

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

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

5 Plusa, B. & Hadjantonakis, A.-K. (De)constructing the blastocyst: Lessons in self-organization from the mouse. *Current opinion in systems biology* **11**, 98-106 (2018). https://doi.org:10.1016/j.coisb.2018.08.002

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

7 Fancsovits, P. *et al.* Early pronuclear breakdown is a good indicator of embryo quality and viability. *Fertility and sterility* **84**, 881-887 (2005). https://doi.org:10.1016/j.fertnstert.2005.03.068

8 Fléchon, J. E. & Kopecný, V. The nature of the 'nucleolus precursor body' in early preimplantation embryos: a review of fine-structure cytochemical, immunocytochemical and autoradiographic data related to nucleolar function. *Zygote* **6**, 183-191 (1998). https://doi.org:10.1017/s0967199498000112

9 Prados, F. J., Debrock, S., Lemmen, J. G. & Agerholm, I. The cleavage stage embryo. *Human reproduction (Oxford)* **27**, i50-i71 (2012). https://doi.org:10.1093/humrep/des224

10 Pereda, J. & Croxatto, H. B. Ultrastructure of a Seven-Cell Human Embryo. *Biology of reproduction* **18**, 481-489 (1978). https://doi.org:10.1095/biolreprod18.3.481

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

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

13 Schrode, N. *et al.* Anatomy of a blastocyst: Cell behaviors driving cell fate choice and morphogenesis in the early mouse embryo: Anatomy of a Blastocyst. *Genesis (New York, N.Y. : 2000)* **51**, 219-233 (2013). https://doi.org:10.1002/dvg.22368

14 Hogan, B. & Tilly, R. In vitro development of inner cell masses isolated immunosurgically from mouse blastocysts. II. Inner cell masses from 3.5- to 4.0-day p.c. blastocysts. *Journal of embryology and experimental morphology* **45**, 107-121 (1978).

15 Strumpf, D. *et al.* Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development (Cambridge)* **132**, 2093-2102 (2005). https://doi.org:10.1242/dev.01801

16 Ralston, A. *et al.* Gata3 regulates trophoblast development downstream of Tead4 and in parallel to Cdx2. *Development (Cambridge)* **137**, 395-403 (2010). https://doi.org:10.1242/dev.038828

17 Nichols, J. *et al.* Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4. *Cell* **95**, 379-391 (1998). https://doi.org:10.1016/S0092-8674(00)81769-9

18 Avilion, A. A. *et al.* Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & development* **17**, 126-140 (2003). https://doi.org:10.1101/gad.224503

19 Chambers, I. *et al.* Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells. *Cell* **113**, 643-655 (2003). https://doi.org:10.1016/S0092-8674(03)00392-1

20 Mitsui, K. *et al.* The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. *Cell* **113**, 631-642 (2003). https://doi.org:10.1016/S0092-8674(03)00393-3

21 Anani, S., Bhat, S., Honma-Yamanaka, N., Krawchuk, D. & Yamanaka, Y. Initiation of Hippo signaling is linked to polarity rather than to cell position in the pre-implantation mouse embryo. *Development (Cambridge)* **141**, 2813-2824 (2014). https://doi.org:10.1242/dev.107276

22 Hirate, Y., Cockburn, K., Rossant, J. & Sasaki, H. Tead4 is constitutively nuclear, while nuclear vs. cytoplasmic Yap distribution is regulated in preimplantation mouse embryos. *Proceedings of the National Academy of Sciences - PNAS* **109**, E3389-3390; author reply E3391-3382 (2012). https://doi.org:10.1073/pnas.1211810109

23 Nishioka, N. *et al.* The Hippo Signaling Pathway Components Lats and Yap Pattern Tead4 Activity to Distinguish Mouse Trophectoderm from Inner Cell Mass. *Developmental cell* **16**, 398-410 (2009). https://doi.org:10.1016/j.devcel.2009.02.003

24 Nishioka, N. *et al.* Tead4 is required for specification of trophectoderm in pre-implantation mouse embryos. *Mechanisms of development* **125**, 270-283 (2008). https://doi.org:10.1016/j.mod.2007.11.002

25 Yagi, R. *et al.* Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development. *Development (Cambridge)* **134**, 3827-3836 (2007). https://doi.org:10.1242/dev.010223

26 Hirate, Y. *et al.* Polarity-Dependent Distribution of Angiomotin Localizes Hippo Signaling in Preimplantation Embryos. *Current biology* **23**, 1181-1194 (2013). https://doi.org:10.1016/j.cub.2013.05.014

27 Sasaki, H. Roles and regulations of Hippo signaling during preimplantation mouse development. *Development, growth & differentiation* **59**, 12-20 (2017). https://doi.org:10.1111/dgd.12335

28 Frum, T., Murphy, T. M. & Ralston, A. HIPPO signaling resolves embryonic cell fate conflicts during establishment of pluripotency in vivo. *eLife* **7** (2018). https://doi.org:10.7554/eLife.42298

29 Frum, T., Watts, J. L. & Ralston, A. TEAD4, YAP1 and WWTR1 prevent the premature onset of pluripotency prior to the 16-cell stage. *Development (Cambridge)* **146** (2019). https://doi.org:10.1242/dev.179861

30 Dickson, A. D. The form of the mouse blastocyst. *Journal of anatomy* **100**, 335-348 (1966).

31 Gardner, R. L. & Rossant, J. Investigation of the fate of 4-5 day post-coitum mouse inner cell mass cells by blastocyst injection. *Journal of embryology and experimental morphology* **52**, 141-152 (1979).

32 Chazaud, C., Yamanaka, Y., Pawson, T. & Rossant, J. Early Lineage Segregation between Epiblast and Primitive Endoderm in Mouse Blastocysts through the Grb2-MAPK Pathway. *Developmental cell* **10**, 615-624 (2006). https://doi.org:10.1016/j.devcel.2006.02.020

33 Arceci, R. J., King, A. A. J., Simon, M. C., Orkin, S. H. & Wilson, D. B. Mouse GATA-4: a retinoic acid-inducible GATA-binding transcription factor expressed in endodermally derived tissues and heart. *Molecular and Cellular Biology* **13**, 2235-2246 (1993). https://doi.org:10.1128/MCB.13.4.2235

34 Plusa, B., Piliszek, A., Frankenberg, S., Artus, J. & Hadjantonakis, A.-K. Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development (Cambridge)* **135**, 3081-3091 (2008). https://doi.org:10.1242/dev.021519

35 Cockburn, K. & Rossant, J. Making the blastocyst: Lessons from the mouse. *The Journal of clinical investigation* **120**, 995-1003 (2010). https://doi.org:10.1172/JCI41229

36 Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev Biol* **375**, 54-64 (2013). https://doi.org:10.1016/j.ydbio.2012.12.008

37 Krivega, M., Essahib, W. & Van de Velde, H. WNT3 and membrane-associated β-catenin regulate trophectoderm lineage differentiation in human blastocysts. *Molecular human reproduction* **21**, 711-722 (2015). https://doi.org:10.1093/molehr/gav036

38 Wei, Q. *et al.* Bovine lineage specification revealed by single-cell gene expression analysis from zygote to blastocyst. *Biology of reproduction* **97**, 5-17 (2017). https://doi.org:10.1093/biolre/iox071

39 Authority, H. F. a. E. (2019).

40 Nagy, Z. P., Varghese, A. C. & Agarwal, A. *In Vitro Fertilization : a Textbook of Current and Emerging Methods and Devices*. 2nd edn, (Springer, 2019).

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

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

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

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

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

46 Luke, B. *et al.* Likelihood of success at each stage of IVF treatment by maternal age and oocyte source: analysis of the 2004-13 cycles in the SART CORS. *Fertility and sterility* **108**, e346-e347 (2017). https://doi.org:10.1016/j.fertnstert.2017.07.1016

47 Muasher, S., Wilkes, C., Garcia, J., Rosenwaks, Z. & Jones, H. BENEFITS AND RISKS OF MULTIPLE TRANSFER WITH IN VITRO FERTILISATION. *The Lancet (British edition)* **323**, 570-570 (1984). https://doi.org:10.1016/S0140-6736(84)90974-7

48 Lee, A. M., Connell, M. T., Csokmay, J. M. & Styer, A. K. Elective single embryo transfer- the power of one. *Contraception and reproductive medicine* **1**, 11-11 (2016). https://doi.org:10.1186/s40834-016-0023-4

49 Elster, N. Less is more: the risks of multiple births. *Fertility and sterility* **74**, 617-623 (2000). https://doi.org:10.1016/S0015-0282(00)00713-5

50 Callahan, T. L. *et al.* The Economic Impact of Multiple-Gestation Pregnancies and the Contribution of Assisted-Reproduction Techniques to Their Incidence. *The New England journal of medicine* **331**, 244-249 (1994). https://doi.org:10.1056/NEJM199407283310407

51 Jones, G. M. *et al.* Novel strategy with potential to identify developmentally competent IVF blastocysts. *Human reproduction (Oxford)* **23**, 1748-1759 (2008). https://doi.org:10.1093/humrep/den123

52 Cimadomo, D. *et al.* The Impact of Biopsy on Human Embryo Developmental Potential during Preimplantation Genetic Diagnosis. *BioMed research international* **2016**, 7193075-7193010 (2016). https://doi.org:10.1155/2016/7193075

53 Botros, L., Sakkas, D. & Seli, E. Metabolomics and its application for non-invasive embryo assessment in IVF. *Molecular human reproduction* **14**, 679-690 (2008). https://doi.org:10.1093/molehr/gan066

54 Omidi, M., Faramarzi, A., Agharahimi, A. & Khalili, M. A. Noninvasive imaging systems for gametes and embryo selection in IVF programs: a review: NONINVASIVE IMAGING SYSTEMS. *Journal of microscopy (Oxford)* **267**, 253-264 (2017). https://doi.org:10.1111/jmi.12573

55 Racowsky, C., Kovacs, P. & Martins, W. P. A critical appraisal of time-lapse imaging for embryo selection: where are we and where do we need to go? *Journal of Assisted Reproduction and Genetics* **32**, 1025-1030 (2015). https://doi.org:10.1007/s10815-015-0510-6

56 Gardner, D. K., Sakkas, D., Seli, E. & Wells, D. *Human Gametes and Preimplantation Embryos : Assessment and Diagnosis*. (Springer New York, 2013).

57 Gardner, D. K., Lane, M., Stevens, J., Schlenker, T. & Schoolcraft, W. B. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertility and sterility* **73**, 1155-1158 (2000). https://doi.org:10.1016/S0015-0282(00)00518-5

58 Armstrong, S. *et al.* Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane database of systematic reviews* **5**, CD011320-CD011320 (2019). https://doi.org:10.1002/14651858.CD011320.pub4

59 Cota, B. M. & Allen, P. J. The developmental origins of health and disease hypothesis. *Pediatr Nurs* **36**, 157-167 (2010).

60 Nasiri, N. & Eftekhari-Yazdi, P. An overview of the available methods for morphological scoring of pre-Implantation embryos in in vitro fertilization. *Cell journal (Yakhteh)* **16**, 392-405 (2015).

61 Barberet, J. *et al.* Can novel early non-invasive biomarkers of embryo quality be identified with time-lapse imaging to predict live birth? *Human reproduction (Oxford)* **34**, 1439-1449 (2019). https://doi.org:10.1093/humrep/dez085

62 Otsuki, J. *et al.* Noninvasive embryo selection: kinetic analysis of female and male pronuclear development to predict embryo quality and potential to produce live birth. *Fertility and sterility* **112**, 874-881 (2019). https://doi.org:10.1016/j.fertnstert.2019.07.015

63 Manor, D., Drugan, A., Stein, D., Pillar, M. & Itskovitz-Eldor, J. Unequal pronuclear size : A powerful predictor of embryonic chromosome anomalies. *Journal of assisted reproduction and genetics* **16**, 385-389 (1999). https://doi.org:10.1023/A:1020550115345

64 Nagy, Z. P. *et al.* Pronuclear morphology evaluation with subsequent evaluation of embryo morphology significantly increases implantation rates. *Fertility and sterility* **80**, 67-74 (2003). https://doi.org:10.1016/S0015-0282(03)00569-7

65 Sadowy, S., Tomkin, G., Munné, S., Ferrara-Congedo, T. & Cohen, J. Impaired development of zygotes with uneven pronuclear size. *Zygote (Cambridge)* **6**, 137-141 (1998). https://doi.org:10.1017/S0967199498000057

66 Ebner, T. *et al.* Presence, but not type or degree of extension, of a cytoplasmic halo has a significant influence on preimplantation development and implantation behaviour. *Human reproduction (Oxford)* **18**, 2406-2412 (2003). https://doi.org:10.1093/humrep/deg452

67 Ebner, T. *et al.* Occurrence and developmental consequences of vacuoles throughout preimplantation development. *Fertility and sterility* **83**, 1635-1640 (2005). https://doi.org:10.1016/j.fertnstert.2005.02.009

68 Gabrielsen, A., Bhatnager, P. R., Petersen, K. & Lindenberg, S. Influence of zona pellucida thickness of human embryos on clinical pregnancy outcome following in vitro fertilization treatment. *Journal of assisted reproduction and genetics* **17**, 323-328 (2000). https://doi.org:10.1023/A:1009453011321

69 Cohen, J., Wiemer, K. E. & Wright, G. Prognostic value of morphologic characteristics of cryopreserved embryos: a study using videocinematography. *Fertility and sterility* **49**, 827-834 (1988). https://doi.org:10.1016/S0015-0282(16)59892-6

70 Sun, Y. P., Xu, Y., Cao, T., Su, Y. C. & Guo, Y. H. Zona pellucida thickness and clinical pregnancy outcome following in vitro fertilization. *Int J Gynaecol Obstet* **89**, 258-262 (2005). https://doi.org:10.1016/j.ijgo.2005.02.012

71 Lewis, E. I. *et al.* Use of imaging software for assessment of the associations among zona pellucida thickness variation, assisted hatching, and implantation of day 3 embryos. *Journal of assisted reproduction and genetics* **34**, 1261-1269 (2017). https://doi.org:10.1007/s10815-017-0978-3

72 Pera, R. A. R. *et al.* Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nature biotechnology* **28**, 1115-1121 (2010). https://doi.org:10.1038/nbt.1686

73 Hardarson, T., Hanson, C., Sjögren, A. & Lundin, K. Human embryos with unevenly sized blastomeres have lower pregnancy and implantation rates: indications for aneuploidy and multinucleation. *Human reproduction (Oxford)* **16**, 313-318 (2001). https://doi.org:10.1093/humrep/16.2.313

74 Van Royen, E. *et al.* Calculating the implantation potential of day 3 embryos in women younger than 38 years of age: a new model. *Human reproduction (Oxford)* **16**, 326-332 (2001). https://doi.org:10.1093/humrep/16.2.326

75 Scott, L., Finn, A., O'Leary, T., McLellan, S. & Hill, J. Morphologic parameters of early cleavage-stage embryos that correlate with fetal development and delivery: prospective and applied data for increased pregnancy rates. *Human reproduction (Oxford)* **22**, 230-240 (2007). https://doi.org:10.1093/humrep/del358

76 Balaban, B. *et al.* Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Reproductive biomedicine online* **22**, 632-646 (2011). https://doi.org:10.1016/j.rbmo.2011.02.001

77 Le Cruguel, S. *et al.* Early compaction at day 3 may be a useful additional criterion for embryo transfer. *Journal of assisted reproduction and genetics* **30**, 683-690 (2013). https://doi.org:10.1007/s10815-013-9983-3

78 Desai, N. N., Goldstein, J., Rowland, D. Y. & Goldfarb, J. M. Morphological evaluation of human embryos and derivation of an embryo quality scoring system specific for day 3 embryos: a preliminary study. *Human reproduction (Oxford)* **15**, 2190-2196 (2000). https://doi.org:10.1093/humrep/15.10.2190

79 Skiadas, C. C., Jackson, K. V. & Racowsky, C. Early compaction on day 3 may be associated with increased implantation potential. *Fertility and sterility* **86**, 1386-1391 (2006). https://doi.org:10.1016/j.fertnstert.2006.03.051

80 Ebner, T. *et al.* Morphological analysis at compacting stage is a valuable prognostic tool for ICSI patients. *Reproductive biomedicine online* **18**, 61-66 (2009). https://doi.org:10.1016/S1472-6483(10)60425-7

81 Ivec, M. B. S., Kovacic, B. P. D. & Vlaisavljevic, V. P. D. Prediction of human blastocyst development from morulas with delayed and/or incomplete compaction. *Fertility and sterility* **96**, 1473-1478.e1472 (2011). https://doi.org:10.1016/j.fertnstert.2011.09.015

82 Maître, J.-L., Niwayama, R., Turlier, H., Nédélec, F. & Hiiragi, T. Pulsatile cell-autonomous contractility drives compaction in the mouse embryo. *Nature cell biology* **17**, 849-855 (2015). https://doi.org:10.1038/ncb3185

83 Marcos, J. *et al.* Collapse of blastocysts is strongly related to lower implantation success: a time-lapse study. *Human reproduction (Oxford)* **30**, 2501-2508 (2015). https://doi.org:10.1093/humrep/dev216

84 Viñals Gonzalez, X. *et al.* Contraction behaviour reduces embryo competence in high-quality euploid blastocysts. *Journal of assisted reproduction and genetics* **35**, 1509-1517 (2018). https://doi.org:10.1007/s10815-018-1246-x

85 Le Bin, G. C. *et al.* Oct4 is required for lineage priming in the developing inner cell mass of the mouse blastocyst. *Development (Cambridge)* **141**, 1001-1010 (2014). https://doi.org:10.1242/dev.096875

86 Orr, B., Godek, K. M. & Compton, D. Aneuploidy. *Current biology* **25**, R538-R542 (2015). https://doi.org:10.1016/j.cub.2015.05.010

87 Alberts, B. *et al.* *Essential cell biology*. Fifth edition, international student edition. edn, (W.W. Norton & Company, 2019).

88 Lee, A. & Kiessling, A. A. Early human embryos are naturally aneuploid—can that be corrected? *Journal of assisted reproduction and genetics* **34**, 15-21 (2017). https://doi.org:10.1007/s10815-016-0845-7

89 Bolton, H. *et al.* Mouse model of chromosome mosaicism reveals lineage-specific depletion of aneuploid cells and normal developmental potential. *Nature communications* **7**, 11165-11165 (2016). https://doi.org:10.1038/ncomms11165

90 Greco, E., Minasi, M. G. & Fiorentino, F. Healthy Babies after Intrauterine Transfer of Mosaic Aneuploid Blastocysts. *The New England journal of medicine* **373**, 2089-2090 (2015). https://doi.org:10.1056/NEJMc1500421

91 Lintsen, A. M. E., Braat, D. D. M., Habbema, J. D. F., Kremer, J. A. M. & Eijkemans, M. J. C. Can differences in IVF success rates between centres be explained by patient characteristics and sample size? *Human reproduction (Oxford)* **25**, 110-117 (2010). https://doi.org:10.1093/humrep/dep358

92 Baker, V. L. M. D. *et al.* Factors affecting success rates in two concurrent clinical IVF trials: an examination of potential explanations for the difference in pregnancy rates between the United States and Europe. *Fertility and sterility* **94**, 1287-1291 (2010). https://doi.org:10.1016/j.fertnstert.2009.07.1673

93 Yan, J., Wu, K., Tang, R., Ding, L. & Chen, Z.-J. Effect of maternal age on the outcomes of in vitro fertilization and embryo transfer (IVF-ET). *Science China. Life sciences* **55**, 694-698 (2012). https://doi.org:10.1007/s11427-012-4357-0

94 Wu, Y., Kang, X., Zheng, H., Liu, H. & Liu, J. Effect of Paternal Age on Reproductive Outcomes of In Vitro Fertilization. *PloS one* **10**, e0135734-e0135734 (2015). https://doi.org:10.1371/journal.pone.0135734

95 Lasiene, K., Vitkus, A., Valanciūte, A. & Lasys, V. Morphological criteria of oocyte quality. *Medicina (Kaunas)* **45**, 509-515 (2009).

96 Said, T. M. & Land, J. A. Effects of advanced selection methods on sperm quality and ART outcome: a systematic review. *Human reproduction update* **17**, 719-733 (2011). https://doi.org:10.1093/humupd/dmr032

97 Mitchell, T. M. *Machine learning*. (WCB/McGraw-Hill, 1997).

98 Simeone, O. A Very Brief Introduction to Machine Learning With Applications to Communication Systems. *IEEE transactions on cognitive communications and networking* **4**, 648-664 (2018). https://doi.org:10.1109/TCCN.2018.2881442

99 Amini, M.-R. & Usunier, N. *Learning with partially labeled and interdependent data*. (Springer, 2015).

100 Alpaydin, E. *Introduction to machine learning*. Third edition. edn, (The MIT Press, 2014).

101 Dey, A. Vol. 7 (International Journal of Computer Science and

Information Technologies, , 2016).

102 Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory* **28**, 129-137 (1982). https://doi.org:10.1109/TIT.1982.1056489

103 Zhan, Y. Z., Wornyo, D. K., Benuwa, B. B., Ghansah, B. & Banaseka Kataka, F. A Review of Deep Machine Learning. *International Journal of Engineering Research in Africa* **24**, 124-136 (2016). https://doi.org:10.4028/www.scientific.net/JERA.24.124

104 Lee, J.-G. *et al.* Deep learning in medical imaging: General overview. *Korean journal of radiology* **18**, 570-584 (2017). https://doi.org:10.3348/kjr.2017.18.4.570

105 Izadpanahkakhk, M., Razavi, S. M., Taghipour-Gorjikolaie, M., Zahiri, S. H. & Uncini, A. Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning. *Applied sciences* **8**, 1210 (2018). https://doi.org:10.3390/app8071210

106 Michelucci, U. *Advanced applied deep learning : convolutional neural networks and object detection*. (Apress, 2019).

107 Deng, J. *et al.* 248-255 (IEEE).

108 Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE transactions on knowledge and data engineering* **22**, 1345-1359 (2010). https://doi.org:10.1109/TKDE.2009.191

109 Kauffman, S., Rosset, S. & Perlich, C. *Leakage in Data Mining: Formulation, Detection, and Avoidance*. Vol. 6(4) (DBLP, 2011).

110 Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (2016).

111 Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559-572 (1901).

112 Kong, X., Hu, C. & Duan, Z. *Principal component analysis networks and algorithms*. (Springer, 2017).

113 Michelucci, U. (Apress L. P, 2022).

114 Beysolow Ii, T. *Introduction to Deep Learning Using R : A Step-by-Step Guide to Learning and Implementing Deep Learning Models Using R*. (Apress, 2017).

115 Nasraoui, O. & Ben N'Cir, C.-E. *Unsupervised and Semi-Supervised Learning* 73-89 (Springer International Publishing AG, 2018).

116 Maaten, L. & Hinton, G. 2579-2605 (Journal of machine learning research, 2008).

117 Hinton, G. E. (ddfs, 20000).

118 Fenton, J. J. *et al.* Influence of Computer-Aided Detection on Performance of Screening Mammography. *The New England journal of medicine* **356**, 1399-1409 (2007). https://doi.org:10.1056/NEJMoa066099

119 Hua, K.-L., Hsu, C.-H., Hidayati, S. C., Cheng, W.-H. & Chen, Y.-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy* **8**, 2015-2022 (2015). https://doi.org:10.2147/OTT.S80733

120 Cheng, J.-Z. *et al.* Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific reports* **6**, 24454-24454 (2016). https://doi.org:10.1038/srep24454

121 Kallenberg, M. *et al.* Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE transactions on medical imaging* **35**, 1322-1331 (2016). https://doi.org:10.1109/TMI.2016.2532122

122 Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A. & Rueckert, D. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage (Orlando, Fla.)* **65**, 167-175 (2013). https://doi.org:10.1016/j.neuroimage.2012.09.065

123 Suk, H.-I. & Shen, D. 2 edn 583-590 (Springer Berlin Heidelberg).

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

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

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

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

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

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

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

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

132 Fitz, V. W. *et al.* Should there be an "AI" in TEAM? Embryologists selection of high implantation potential embryos improves with the aid of an artificial intelligence algorithm. *J Assist Reprod Genet* **38**, 2663-2670 (2021). https://doi.org:10.1007/s10815-021-02318-7

133 VerMilyea, M. *et al.* Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Hum Reprod* **35**, 770-784 (2020). https://doi.org:10.1093/humrep/deaa013

134 Diakiw, S. M. *et al.* Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. *Human reproduction (Oxford)* **37**, 1746-1759 (2022). https://doi.org:10.1093/humrep/deac131

135 Payá, E., Bori, L., Colomer, A., Meseguer, M. & Naranjo, V. Automatic characterization of human embryos at day 4 post-insemination from time-lapse imaging using supervised contrastive learning and inductive transfer learning techniques. *Computer methods and programs in biomedicine* **221**, 106895-106895 (2022). https://doi.org:10.1016/j.cmpb.2022.106895

136 Sawada, Y. *et al.* Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. *Reprod Biomed Online* **43**, 843-852 (2021). https://doi.org:10.1016/j.rbmo.2021.05.002

137 Ueno, S. *et al.* Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: a single-center large cohort retrospective study. *Fertil Steril* **116**, 1172-1180 (2021). https://doi.org:10.1016/j.fertnstert.2021.06.001

138 Miyagi, Y., Habara, T., Hirata, R. & Hayashi, N. Feasibility of predicting live birth by combining conventional embryo evaluation with artificial intelligence applied to a blastocyst image in patients classified by age. *Reprod Med Biol* **18**, 344-356 (2019). https://doi.org:10.1002/rmb2.12284

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

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

141 Quaas, A. & Dokras, A. Diagnosis and treatment of unexplained infertility. *Reviews in obstetrics and gynecology* **1**, 69-76 (2008).

142 Webster, M., Witkin, K. L. & Cohen-Fix, O. Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *Journal of cell science* **122**, 1477-1486 (2009). https://doi.org:10.1242/jcs.037333