**Universidade de Aveiro** Departamento de Biologia
**2019**

**Oldair E. B. Mendes**

# Embryo selection through time-lapse image analysis: a Deep Learning approach

# Seleção de embriões pela análise de imagens: uma abordagem Deep Learning

# DECLARAÇÃO

Declaro que este relatório é integralmente da minha autoria, estando devidamente referenciadas as fontes e obras consultadas, bem como identificadas de modo claro as citações dessas obras. Não contém, por isso, qualquer tipo de plágio quer de textos publicados, qualquer que seja o meio dessa publicação, incluindo meios eletrónicos, quer de trabalhos académicos.

**Universidade de Aveiro**
**2019**

Departamento de Biologia

**Oldair E. B. Mendes**

# Embryo selection through time-lapse image analysis: a Deep Learning approach

# Seleção de embriões pela análise de imagens: uma abordagem Deep Learning

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biologia Molecular e Celular, realizada sob a orientação científica da Doutora Margarida Fardilha, Professora Auxiliar com Agregação do Departamento de Ciências Médicas da Universidade de Aveiro e do Doutor António Barros, Investigador Auxiliar Convidado da Faculdade de Medicina da Universidade do Porto

Para Anabela

**O júri**

| | |
|---|---|
| Presidente | Prof. Doutor António José de Brito Fonseca Mendes Calado<br>Professor Auxiliar da Universidade de Aveiro |
| Arguente Principal | Doutor Miguel Monsanto Pinheiro<br>Investigador Doutorado da Universidade de Aveiro |
| Vogal- orientador | Doutor António de Sousa Barros<br>Investigador Auxiliar Convidado da Faculdade de Medicina da Universidade do Porto |

**Agradecimentos**

Obrigado,

Anabela pelo apoio incondicional e constante, e à sua família pela compreensão e apoio;

Prof. Doutora Margarida Fardilha e Doutor António Barros, por acreditarem que podia levar avante este projeto, pela orientação, disponibilidade e independência que me proporcionaram e por tornar menos sinuoso este percurso;

Clínica Ferticentro e seu embriologista clínico Elisson Souza, pela prontidão na disponibilização, organização e anotação das imagens usadas neste projeto;

Caros colegas pelo acolhimento e por dar-me a conhecer esta magnífica cidade e nobre universidade;

Minha mãe, o exemplo que me guia.

A todos muito obrigado!

**palavras-chave** Machine Learning, IVF, Imagens Time-lapse, Seleção de embriões, Deep Learning

**Resumo**

A infertilidade afeta cerca de 186 milhões de pessoas em todo o mundo e 9-10% dos casais em Portugal, causando problemas financeiros, sociais e de saúde. Constitui procedimento padrão a avaliação da qualidade dos embriões baseadas em características morfológicas. No entanto, tais avaliações são subjetivas e demoradas e resultam em classificações discrepantes entre embriologistas e clínicas causando problemas na avaliação do potencial do embrião. Embora as tecnologias de reprodução medicamente assistida, como a fertilização *in vitro*, acoplada à tecnologia time-lapse, tenham diminuído o problema da infertilidade, existem limitações significativas, mesmo considerando a análise morfocinética. Outrossim, muitas pacientes necessitam de múltiplos ciclos de fertilização para alcançar a gravidez, tornando a seleção do embrião com maior potencial de implantação e geração de nados vivos um desafio crítico. No presente projeto demonstramos a prova do conceito da confiabilidade de Machine Learning (aprendizagem automática), especialmente Deep Learning baseado em TensorFlow e Keras, para extrair e discriminar caraterísticas associadas ao potencial embrionário, em imagens time-lapse. Igualmente, apresentamos um pipeline para que clínicos e investigadores, sem experiência em Machine Learning, possam utilizar com facilidade, rapidez e precisão Deep Learning como ferramenta de apoio à decisão clínica em estudos de viabilidade de embriões, bem como noutras áreas médicas onde a análise de imagens seja proeminente.

**Keywords** Machine Learning, IVF, Time-lapse image, Embryo Selection, Deep Learning

**Abstract**

Infertility affects about 186 million people worldwide and 9-10% of couples in Portugal, causing financial, social and medical problems. Evaluation of embryo quality based morphological features is the standard *in vitro* fertilization (IVF) clinics around the world. This process is subjective and time-consuming, and results in discrepant classifications among embryologists and clinics, leading to fail in predict accurately embryo implantation and live birth potential. Although assisted reproductive technologies (ART) such as IVF coupled with time lapse elimination of periodic transfer to microscopy assessment and stable embryo culture conditions for embryos development, has alleviated the infertility problem, there are significant limitations even considering morphokinetic analysis. Likewise, many patients require multiple IVF cycles to achieve pregnancy, making the selection of single embryo for transfer a critical challenge. Here, we demonstrate the reliability of machine learning, especially deep learning based on TensorFlow open source and Keras libraries for embryo raw TLI images features extraction and classification in clinical practice. Equally, we present a follow up pipeline for clinicians and researchers, with no expertise in machine learning, to easily, rapid and accurately utilize deep learning as a clinical decision support tool in embryos viability studies, as well in other medical field where the analysis of images is preeminent.

## Contents

## Lists of Figures

## List of Tables

## List of Abbreviatures **(in order of appearance)**

| | |
|---|---|
| **IVF** | *In Vitro* Fertilization |
| **ART** | Assisted Reproductive Technique |
| **TLI** | Time Lapse Imaging |
| **ESHERE** | European Society of Human Reproduction and Embryology |
| **ICM** | Inner Cell Mass |
| **OCM** | Outer Cell Mass |
| **TE** | Trophectoderm |
| **ZP** | Zona Pellucida |
| **SET** | Single-Embryo Transfer |
| **NPB** | Nucleolar Precursor Bodies |
| **TLS** | Time-Lapse Systems |
| **ICSI** | Intracytoplasmic Sperm Injection |
| **TLM** | Time-Lapse Monitoring |
| **Eeva** | Early Embryonic Viability Assessment |
| **KID** | Known Implantation Data |
| **AI** | Artificial Intelligence |
| **ML** | Machine Learning |
| **DL** | Deep Learning |
| **ANN** | Artificial Neural Network |
| **CNN** | Convolutional Neural Network |
| **ReLu** | Rectified Linear Unit |
| **CPU** | Central Processing Unit |
| **GPU** | Graphics Processing Unit |
| **API** | Application Programming Interface |
| **AUC** | Area Under Curve |
| **ROC** | Receiver Operating Characteristics |
| **TPR** | True Positive Rate |
| **FPR** | False Positive Rate |
| **TP** | True Positive |
| **TN** | True Negative |
| **FP** | False Positive |
| **FN** | False Negative |

# I. INTRODUCTION

## 1. PREAMBLE

Infertility is defined as the failure to conceive after 12 months of regular, unprotected sexual intercourse (six months if the women is over 35 years) or the inability to maintain the pregnancy until term[1]. It affects about 186 million people worldwide[2] and 9-10% of couples (260 to 290 thousand couples) in Portugal, according to Portuguese Infertility Association.

Since the birth of first assisted conceived baby, *in vitro* fertilization (IVF) has become the most common assisted reproductive technique (ART) for infertility treatment. IVF involves ovarian simulation followed by retrieval of several oocytes, fertilization and embryo culture for 1-6 days, under a restricted laboratory condition. The final goal is a birth of single healthy and normal baby. To increase pregnancy rate multiple embryo transfer are normally achieved, elevating the risk of multiple gestations, increasing as well financial, social and medical implications. This trade-off between successful outcomes and multiple pregnancy lead most countries to adopt more rigids policies in IVF. Selection of a single embryo with right developmental competence become the practices among the ART´s. Consequently, visual quality assessment of an embryo´s shape and development are main determinant of implantation and pregnancy in IVF. While this image selection method is universal in clinical practice, is subjective and time consuming. With the introduction of time-lapse imaging (TLI) this morphological assessment has become more objective and allows the identification of the very best embryo for uterine transfer.

Henceforth, 40-60% of good quality time-lapsed embryo in IVF fail to conduct to pregnancy and only 4.7% result in live births[3]. Indeed, these low performant results could and should improve if one use complementary probing approaches. There is a close correlation between morphological parameters and stages of development of the embryo at given time points and developmental competence such as defined in Alpha[a] and European Society of Human Reproduction and Embryology (ESHERE) consensus[4], but is dependent and subject to inter and intra-operator variations. Machine learning oriented clinical decision, coupling time-lapse imaging and deep learning could potentially improve this long-standing problem.

[a] International forum for scientists in reproductive medicine. For more info: http://alphascientists.org

Following Payne and co-workers first publication to report *in vitro* monitorization of embryo developmental events[5], many time-lapse based algorithms have been developed trying to investigate whether morphological and kinetics markers can assist in embryo selection and predict overall implantation rate. Therefore, many of these morphokinetics algorithms are study-specific, that is, they are mostly performant when applied to the clinics own data[6], denoting clearly a biased orientation; failing when applied in a cross study. Compared to conventional embryo incubation, there is no clear differences in clinical pregnancy, live birth or stillbirth[7], indeed there is a lack of robust and fully automated method to analyse TLI data.

## 2. HUMAN *IN VITRO* FERTILIZATION

Globally more than five million babies have been born from assisted conception and this population is now increasing by over one million *per annum* - of which a large number was by in vitro fertilization (IVF)[8]. Although the procedure is not fully established until the last decades of the 20th century, the history of IVF goes back to several decades early. It was first studied in non-mammalian species. The first observation of sperm penetration into an egg was reported in *Ascaris mystax* by Nelson in 1851[9]. Since then, studies in non-mammalian species have provide crucial details of the fertilization process.

The first IVF on mammalian eggs was performed by S.L. Schenk, in 1878, working with rabbit and guinea pig, where he noted that cell division happens in culture after sperm were added to ova[10]. In 1935 Gregory Pincus, described the first experiment that allowed rabbit oocytes to mature *in vitro*, reaching the metaphase stage of meiosis II[11]. At same United State research institute as Pincus - Worcester Foundation for Experimental Biology, Chang in 1959 showed that *in vitro*-matured rabbit oocytes could be fertilized *in vitro* and give rise to viable embryos. When transferred back to adult females, they produce viable offspring[12]. Chang's findings represented a significant advance. Still there was a need for pre-incubation of the sperm in the female uterus prior to the attempt to fertilize the oocyte[12]. In 1963, Chang and Yanagimachi correct the initial Chang idea, when they identified experimental *in vitro* conditions through which hamster spermatozoa could fertilize oocytes, without prior *in vivo* sperm activation, and give rise to 2-cell stage embryos[13].

At beginning of the 20$^{th}$ century, reproduction researchers started to discuss the possibility and conditions that could allow the fertilization of human oocytes, with no progress until early 1960s, due in large part to the complexity of fertilization processes, despite the significant advances in animal research. The problems faced by the early human reproduction investigators include the control the oocyte maturation process, retrieve oocytes at a developmental stage suitable for IVF, the ability to activate sperm *in vitro*, define conditions that would promote fertilization as well as early embryo development *in vitro* and finally, a method through which early embryos could be transferred back to the uterus of the mother[14–16].

In 1965, Robert G. Edwards solved the problems of the access of mature oocytes for IVF and identify buffer conditions to promote their maturation *in vitro*, finding that they require 24 hours of incubation before initiating the maturation process[14,15,17]. In the final years of the 1960s Steptoe publish a laparoscopy method that allowed the visualization of the human female reproductive tract and opening the possibility to aspirate oocytes from the ovary[18,19].

The combination of Edwards and Steptoe findings would be crucial to the born of the first healthy IVF conceived baby, named Louise J. Brown in 1978, as a result of transferring early embryos from IVF back into female uterus. The Edwards´s IVF process is shown in Figure 1.

Human IVF is a process of insemination of female oocytes by sperm male cells, involves ovarian stimulation followed by retrieval of several oocytes, fertilization and embryo culture for 2-3 days (cleavage stage) or 5-6 days (blastocyst stage), under a restricted laboratory condition. Once embryos are formed, they are placed in the uterus. It was initially adopted for treatment of female with inoperable tubal blockage. Since 1978, when the first *in vitro* conceived normal healthy baby was born[16,20] IVF has evolved considerably to become the most important ART to treat infertility worldwide[21]. Today IVF is indicated for treatment of absence of functional fallopian tube, endometriosis, low sperm counts and/or quality and refractory anovulation[1,2].

IVF born children are, in general, as healthy as children born after natural conception. However, is denoted a higher prevalence of multiple births associated with IVF treatments compared to normal pregnancies[8,22–26], largely due to multiple embryo transfer during IVF cycles. Multiple embryo transfer results in multiple pregnancies and associated perinatal and postnatal health complications to the mother as well to the babies. To avoid the consequences of multiple pregnancy many countries adopted restrictive policies limiting the number of embryos to be transferred in a single cycle, adding a challenge to select only the most viable ones for transfer[1,8]. Making the identification of embryo with greatest potential to develop a foetus the main problems faced by this treatment.

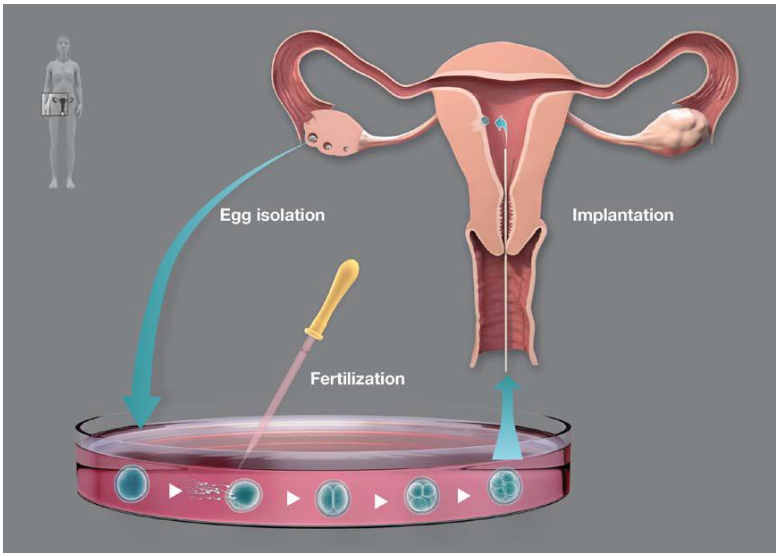



**Figure 1: Edwards´s IVF process[b].** By Steptoe´s Laparoscopy oocytes are retrieved prior to fertilization and placed in a culture dish with medium. Sperm are added and activated by the dish conditions. The fertilization resulting from sperm and egg fusion, form an embryo that undergo a series of cell division, since 8-cell stage when they are transferred back to the uterus, using thin needle. Further embryo development will take place between wall of uterus and endometrium.

### 2.1. EMBRYO GRADING SYSTEMS

Human embryo development begins with fertilization, a process by which the male spermatozoon fuse with female oocyte to give rise to a new organism – the zygote (1-cell)[27]. The single cell divide itself by first mitosis to become a two-cell embryo, at first cleavage about 19 hours after fertilization. The two-cell (2C) divide, during the second and third cleavage to 3-cell and 4-cell embryo on day 2 and so on[28]. These blastomeres cells, become smaller with each cleavage division until 8-cell stage, where they are loosely arranged. At day 4, embryo start to compact into a 16-cell morula and shortly develop a blastocoel – a fluid filled cavity[27]. Blastocoel formation define a blastocyst stage[29]. At day 5, the embryonic cells differentiate and inner cells become the inner cell mass (ICM) and the surround cells the outer cell mass (OCM). The ICM develop into the embryo and OCM into trophectoderm (TE). TE is responsible for the formation and subsequent expansion of the blastocoel which later

[b] Advanced information. NobelPrize.org. Nobel Media AB 2019. Wed. 23 Oct 2019. https://www.nobelprize.org/prizes/medicine/2010/advanced-information/

contribute to the placenta[30]. At day 5-6, usually occurs a hatching, a process by which mammalian embryo escape from the multilayer glycoprotein membrane called zona pellucida (ZP). The early human embryo development is shown in Figure 2.

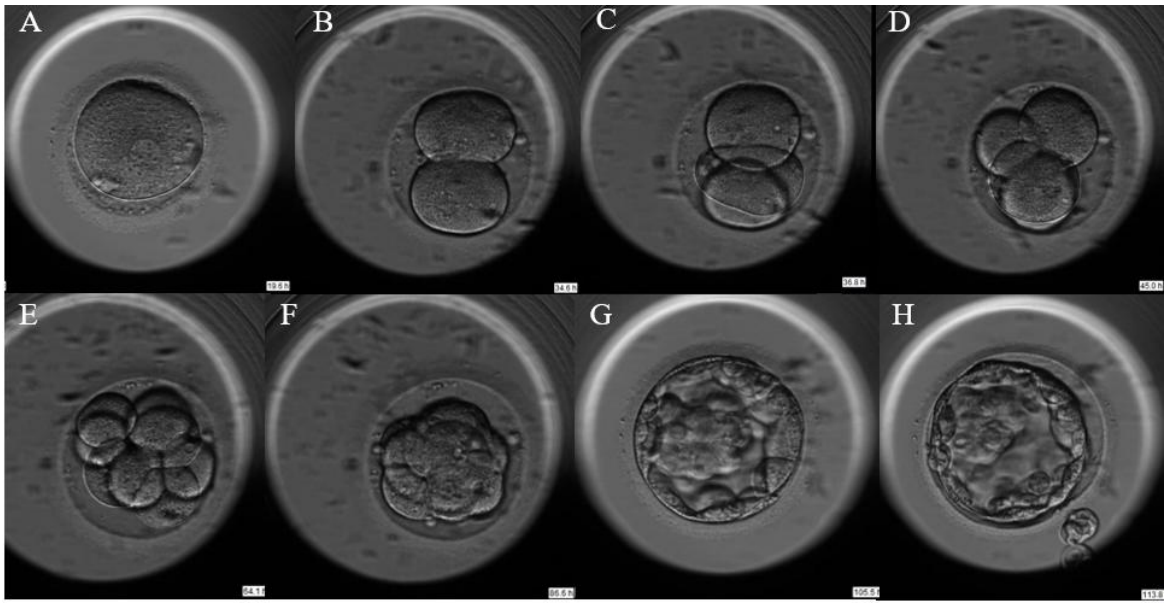



**Figure 2: Outline of early stages of development of pre-implanted human embryo.** (A) oocyte at pronuclear stage at 19 hours; (B) 2-cell stage at 34 hours; (C) 3-cell stage at 37 hours;(D) 4-cell stage at 45 hours; (E) 8-cell stage at 63 hours; (F) morula stage at 87 hours; (G) blastocyst stage at 105 hours; (H) expanding blastocyst at 114 hours, prior embryo transfer. Each embryo develops at different rate. Images kindly provided by Ferticentro, Fertility Study Center, S.A., Coimbra, Portugal.

One of most challenging problem in infertility by IVF treatment is a precise embryo quality evaluation, crystalized in the selection of the embryo with greatest developmental competence. Examination and therefore graduation of embryo resulting in the improved pregnancies outcome and reduced number of embryos to be transferred to the uterus, thus avoiding multifetal pregnancies[31].

Currently practice recommend Single embryo transfer (SET)[32]. SET has a high benefit to the mother as well to the babies[31,33,34]. To select the most viable and competent embryo, elegant grading systems have been developed over time. They vary according to stage of development of the embryo and rely in visual morphology and, with the introduction of time-lapse imaging technology, kinetics analysis[35–37] (see time-lapse section, 2.2).

Most of grading schemes are based in degree of fragmentation, presence and number of nuclei, number and symmetry of blastomeres. Blastocysts are evaluated considering expansion of blastocoel, the number and properties of cells in ICM and TE. Key morphological features

for embryo scoring are shown in Figure 3. However, all different scoring systems can be grouped considering three main stages: zygote, embryo cleavage and blastocyst[38]. With the introduction of TLI, researchers and embryologist focus on the key time of specific developmental events such as pronuclei fading, first cleavage or duration of cell cycle[39].

To reduce inter-laboratory variations and in the quest of an international consensus in the morphological assessment of embryo, in 2011, the Alpha Executive and ESHRE Special Interest Group of Embryology, published the results of a convened workshop to establish common criteria and terminology for grading oocytes, zygotes and embryos that would be used in routine application in any IVF laboratory[4]. The consensus is presented at each scoring systems.

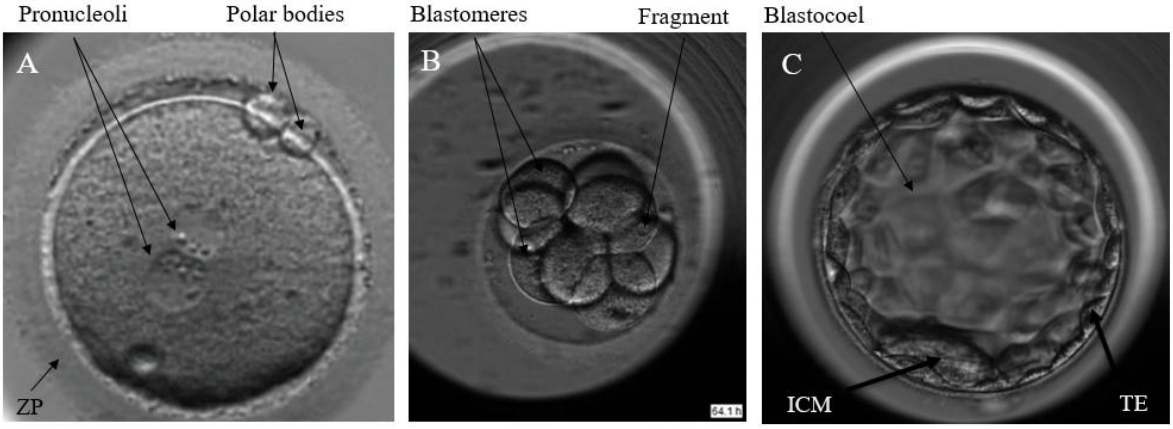



**Figure 3: Key morphological features at different developmental stage**. (A) fertilized oocyte with pronuclei and polar bodies (B) Cleaved embryo showing blastomeres and fragment formation and (C) blastocyst with blastocoel fluid cavity, ICM (Inner Cell Mass) and TE (trophectoderm). ZP stand for Zona Pellucida. Image A adapted from[39], B and C kindly provided by Ferticentro, S.A.

### 2.1.1. PRONUCLEATE OOCYTE SCORING

Normally the zygote evaluation is done about 16-18 hours after fertilization, considering the formation and subsequent size increase of pronuclei, the symmetry, size, number, quality and distribution of nucleoli and the breakdown of pronuclear membrane[39].

Payne and colleagues were the first to document earliest events of embryo development by video recording body extrusion and formation of the male and female pronuclei and relating them with embryo quality[5].

The fact that alignment and number of nucleoli within each nucleus can be used to assess human embryo developmental outcome, was shown by Scott & Smith[40] and Tesarick &

Greco[41], leading Scott and co-workers to develop pronuclear scoring system[42], widely accepted in the reduction of embryo number required to achieve pregnancy[39] and selection of embryos with better implantation rates. Scott scoring consider five grades based on both number and distribution of nucleoli in the pronuclei, been 1 the high quality one and 5 for one with poor developmental potential. Zygote grade 1 to 3 (Z1 to Z3) are proper for IVF (See Table 1, for comparison)[43].

The Alpha Executive and ESHRE Special Interest Group of Embryology consensus[4] consider as optimal oocyte morphology that with a spherical structure enclosed by a uniform zona pellucida, uniform translucent cytoplasm free of inclusions and a size-appropriate polar body. They propose using three categoric systems for grading pronuclei symmetry: category 1 as good quality for symmetrical one, 2 representing medium quality and category 3 for abnormal and lowest quality (Table 1).

In 2014, Aguilar and co-workers investigate morphology of fertilization events such as second polar body extrusion, pronuclear fading and length of S-phase by time-lapse technology, concluding that the timings they occurred, 3.3-10.6 h, 22.2-25.9 h and 5.7-13.8 h, respectively, were linked to successful embryo implantation[44].

Gardner and Balaban analysing morphological scoring systems, such as presented by Aguilar, time-lapse imaging technology and associated algorithms for computer-assisted scoring, propose as key features for high viable oocytes, the following characteristics: number of nucleolar precursor bodies (NPB) in both pronuclei never differed by more than 3; NPB must be always polarized or not polarized in both pronuclei but never differently polarized and the angle from the axis of pronuclei and the furthest polar body must be less than 50°[39].

**Table 1:** Alpha and ESHRE Consensus scoring system for pronuclei. Adapted from**[4].**

| Category | Rating | Description |
|---|---|---|
| 1 | Symmetrical | Equivalent to Z1 and Z2 from Scott scoring (2003) and Gardner & Balaban key features for high viable oocytes |
| 2 | Non-symmetrical | Other arrangements, incluidng peripheral sited pronuclei |
| 3 | Abnormal | Pronuclei with one or no NPB |

### 2.1.2. CLEAVAGE STAGE SCORING

There are many scoring systems based on morphological features of the splitting embryo such as degree of fragmentation, symmetry of the blastomeres, presence of multinucleation or

compaction status[45–50]. The Alpha and ESHRE consensus[4] consider that an optimal day 2 embryo would have four equally sized mononucleated blastomeres in a three-dimensional tetrahedral arrangement, with 10% fragmentation. Day 3 embryo would have 8 equally sized mononucleated blastomeres, with 10% fragmentation (Table 2) at which should be added the cell number.

**Table 2:** Alpha and ESHRE Consensus scoring system for cleavage stage. Adapted from**[4].**

| Category | Rating | Description |
|---|---|---|
| 1 | Good | Less than 10% fragmentation, stage-specific cell size and no multinucleation |
| 2 | Fair | 10-25% of fragmentation, stage-specific cell size for majority of cell and no evidence of multinucleation |
| 3 | Poor | Severe fragmentation (more than 25%), Cell size not stage-specific and evidence of multinucleation |

Gardner and Balaban propose mononucleated blastomeres, equal cell size and less than 20% of fragmentation for early cleavage and adding at least 7 blastomeres for day 3 cleavage as key morphological features of the cleavage stages[39].

### 2.1.3. BLASTOCYST SCORING

Blastocyst stage is characterized by formation of blastocoel at the centre of the embryo, surrounded internally by TE and ICM. Externally is surrounded by ZP until hatching[51]. Refer to Figure 3C.

Gardner and Lane[52] introduced the blastocyst transfer in human IVF in 1997. Since then several grading systems were proposed for embryo blastocyst grading. The commonly used is the Gardner classification[53]. According to this grading system each blastocyst embryo consists in three individual scores. The development of blastocyst ranges from 1 to 6, being 6 for the hatched blastocyst. The ICM is qualified as A, B or C, being A for the tightly packed and many cells ICM. The TE quality are classified as a, b or c, being a for one with many cells that form tightly epithelium. Indeed, Van den Abbeel and co-workers[54] confirm that the degree of expansion of blastocyst, assessment of high quality TE and ICM increase pregnancy and live birth rates. The same authors concluded that blastocyst with ICM grade A may reduce the risk of early pregnancy loss. Du and colleagues[55], using logistic regression analysis, confirmed the importance of blastocyst expansion, as well ICM and TE quality but in different extension, in

the prediction of live birth. Zhao *et al*., in a recent publication re-investigate the expansion of blastocoel as essential for successful pregnancy[56].

For the Alpha consensus an optimal blastocyst stage embryo should be fully expanded, with a distinguished and many cells ICM and cohesive epithelium TE, as shown in Table 3.

**Table 3:** Alpha and ESHRE Consensus scoring system for blastocyst stage. Adapted from[4].

| | Categoty | Rating | Description |
|---|---|---|---|
| **Stage of development** | 1 | | |
| | 2 | | Early blastocyst expanded |
| | 3 | | |
| | 4 | | Hatched/hatching |
| **ICM** | 1 | Good | Proeminent, easily discemible with many cells that are compacted and tighly adhered toghter |
| | 2 | Fair | Easily discemible with many cells that are loosely and grouped toghter |
| | 3 | Poor | Difficult to discern, with few cells |
| **TE** | 1 | Good | Many cells formimg a cohesive epithelium |
| | 2 | Fair | Few cells formimg a loose epithelium |
| | 3 | Poor | Very few cells |

Only about 20-40% of embryo morphologic-based characterization will be implanted[57]. With the advances in TLI many researchers starting to use these images for grading day 5 embryos. Santos Filho[58] presented in 2012, a semi-automated method using Support Vector Machine (SVM) classifier. In 2015, Singh e*t al*.[59] following Santos Filho studies, publish a fully automated method for segmentation and measurement of TE region of blastocyst. Both studies proposed to automate human embryo analysis and eliminate the inter-observer variations from previously grading methods.

### 2.2. AUTOMATION OF HUMAN EMBRYO ANALYSIS

Most of grading systems for human embryo quality assessment are subject to inter- and intra-observer variation. To mitigate this problem computer-assisted scoring and automation of embryo visualization has been shown to improve embryo quality assessment by providing quantitative and objective information to complement traditional morphological analysis.

### 2.2.1. TIME-LAPSE IMAGING TECHNOLOGY

Time-lapsed animal embryos observation were reported as early as the 1950's[60], but had not been applied to human embryos until the late 1990's[5]. Payne and colleagues[5] monitoring *in vitro* embryo events such polar body extrusion and formation of the male and female pronuclei, during 17-20 hours of intracytoplasmic sperm injection (ICSI). The team analysed such fertilization events in 50 oocytes that underwent ICSI in a Perspex chamber equipped with an Olympus IX-70 inverted microscope. The vacant decades from animal to human experiments were due to technical limitations in the manufacture of the incubator and high-resolution automated imaging technology. Early monitoring of human embryo was performed with internal equipment mainly for experimental purposes and observing only one embryo/oocyte at a time[5,61].

Only in 2010, the time-lapse monitoring of human embryos for therapeutic purposes has been emerging in the literature, with the use of commercially available time-lapse devices[62].

Other technologies has been experimented such as preimplantation genetic screening[63,64], metabolomics or proteomics[65] to identify embryo with high implantation potential[35,66], but some of them require the use of complex technology, embryo cryopreservation[35,66] and thus increasing treatment expenses[57].

Over recent years time-lapse systems (TLS), has been presented as alternative to traditional displacement of embryo from conventional incubators for quality assessment. This non-invasive microscopic technology offers solutions to overcome most of the problems of the standard embryo morphological analysis. It allows safely incubation of embryo in stable culture conditions, minimizing the potential impact of changes in temperature or gas composition, enhance our knowledge of embryo kinetics and allow the evaluation of quantitative and qualitative parameters[67], improving single blastocyst selection for SET[3,66] and reducing inter- and intra-observer variations (Figure 4). TLS provides uninterrupted and precise timing events of embryo development such as pronuclear formation, early cleavage, cell cycle intervals, cell division and initiation of blastulation, without removing them from incubators[67]. Aiding the likelihood of selecting the best single embryo for uterine transfer.

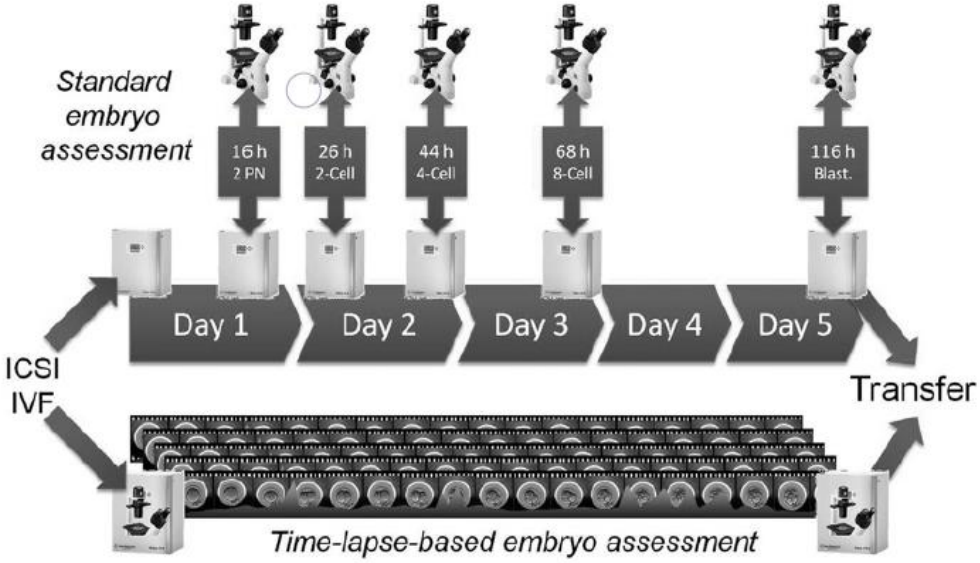



**Figure 4: Time-lapse-based and traditional embryo assessment.** TLM eliminate periodic transfer of the embryos to microscopy for morphological assessment, allowing stable culture conditions for embryos development. The time points for quality assessment are according from Alpha and ESHRE consensus. Adapted from[67].

Nowadays, there are four commercially available time-lapse systems: Primo Vision (Vitrolife), Embryoscope (Fertilitech)[68], Eeva (Early Embryonic Viability Assessment, Auxogyn) system and Ecso Miri (Ecso Medical)[69]. All of them consist on the same technology strategies, a digital inverted microscope camera that takes pictures of embryos at selected intervals. With proper software, a video follows their development. Ecso Miri and Embryoscope both use a compact, self-contained incubator with built-in camera. While, Primo Visio and Eeva use cameras that is placed in a traditional incubator[67]. Each system differs in source of light and in the way that embryo is transported to the field of view. Primo Visio, Embryoscope and Ecso Miri uses bright fields technology allowing the assessment both kinetics and morphologic parameters, instead Eeva uses dark field technology thus allowing limited morphologic parameters[70].

There are many evidences of the technology usefulness, but does the use of TLM improve outcomes for embryo incubation and selection? Several studies have claimed favourable outcomes in prediction of blastocyst formation[36,71,72], implantation[73,74] and pregnancy rate[74]. However as shown by Chen *et al.*[75] in a meta-analysis and systematic review of randomized controlled trials, clinical TLI may have the potential to improve IVF outcomes but currently there is insufficient evidence to support regular TLI use when considering

ongoing pregnancy rates and blastocyst formation rates, that is, there no sufficient evidence to support that TLI is superior to conventional methods for human embryo incubation and selection. The idea are supported by Armstrong and co-workers[7] as well as several other authors[76–78].

High and successful implantation IVF is achieved by transferring embryos with highest developmental competence[79]. Thus, to select just one highly developmental competent embryo, reliable and informative biomarkers is sought.

Payne and other TLM findings encourage many other groups to investigate whether kinetic markers can assist in embryo selection and, thus, predict implantation potential.

### 2.2.2. TIME-LAPSE IMAGING-BASED ALGORITHMS

The first algorithm designed to improve clinical outcomes by predicting embryo viability come from Wong *et al.*[36]. The authors correlate time-lapse image analysis and gene expression profiling through preimplantation development from the zygote to the blastocyst stage, predicting which embryo will reach blastocyst stage using three dynamics parameters: cytokinesis, time between first and second mitoses and time between second and third mitoses.

After the publication of consensus paper on morphological criteria for embryo assessment from Alpha and ESHRE experts[4], late in 2011, and the advances of time-lapse monitoring imaging technology allowing the mapping of morphological changes with exact timepoint[5,35] coupled, in time, to the definition of such criteria[80], several studies have been published describing morphokinetic algorithm based on parameters defined by time-lapse imaging to improve implantation rate.

Meseguer team[81] and Basile *et al.*[82] presented algorithms that select the best embryo for transfer by morphologically screening embryos and assessing them for the presence of exclusion criteria. These algorithms follow different hierarchical classification trees with eight morphokinetic scoring levels ($A^+$ as the highest to F as the lowest). Meseguer classifies embryos through a combination of morphological assessment, inclusion and exclusion criteria. Indeed, this algorithm consider not only morphological criteria but also kinetics markers.

Conaghan and co-workers[83] using data from five United State fertilization clinics, present an algorithm based on early cleavage time intervals, that combine time-lapse image analysis with cell-tracking software, Eeva to measure early embryo development and generate blastocyst predictions by day 3. It includes two categories: Eeva high and Eeva low. Vermilyea *et al.*[84] extend Conahan studies adding Eeva medium to the previously categories. Using data from six clinics and Eeva imaging systems, the authors examine and stablish relationships

between computer outputs derived by Eeva (High, Medium and Low) with embryo implantation and clinical pregnancy. Eeva algorithm could provide valuable information, improve the success of cleavage stage and facilitate the trend of SET.

Goodman *et al.*[85] designed an algorithm based on morphological and kinetics parameters, for day 5 embryo selection. The embryo is select by an accumulation of positives and negatives points, from 4 to -2 points. This algorithm does not improve clinical reproductive outcomes but is associated with blastocyst implantation rate.

In the same year (2016) Liu *et al*.[86] used embryo with known implantation data (KID) at day 3. The algorithm uses quantitative features such as poor conventional day 3 morphology, abnormal cleavage patterns and qualitative ones as pronuclear fading to 5-cell stage and duration of 3-cell stage. Grading them in 7 levels from $A^+$ (highest) to F (lowest), using a morphokinetic algorithm with pronuclear fading as reference starting point.

According to Peterson *et al.*[87] the majority on these algorithms are at risk of only performing adequately on dataset it was developed because they are based on specific data with few been tested in prospective trials. Some of these algorithms[37,82] have been tested on independent datasets or in other clinical settings with varying outcomes. Petersen and colleagues[87] presented a KIDScore algorithm, that like a Liu *et al*. is also based on KID data from 24 clinics, ranking time-lapse monitored embryos according to their blastocyst formation, independently of culture conditions and fertilization method. This algorithm does not require initial screening phase based on morphology and score the embryos only based on decision tree (from 5 to highest to 1 to the lowest). The developed KIDScore algorithm are based on pronuclei number at 1-cell stage and time from insemination to 2-cell, 3-cell, 5-cell and 8-cell stage, and predict implantation potential with area under the curve (AUC) of 0.650, obtaining AUC of 0.745 for blastocyst development and 0.679 for its quality.

Kovacs *et al*.[66] in prospective, randomized and controlled studies carried in two clinics from 2012-2015, determined whether a selection of a single blastocyst based on an algorithm comprising kinetic and morphologic scores assessed through continuous time-lapse monitoring results improve clinical outcome compared to embryo selection based on morphology alone and assess whether a time-lapse score based on kinetic and morphologic parameters was predictive of implantation, concluding that selection of a single blastocyst based on information derived from time-lapse monitoring can aid embryo selection for single embryo transfer.

Storr *et al*.[6] (2018) in a prospective study examined the agreement among some of these algorithms and between then and embryologists, found that is highly variable and may be site-

specific and they involve low number of embryos from single center and lack of validation in independent studies. Otherwise, in recent studies, Armstrong and colleagues[7], aimed to determine the effect of a time-lapse systems (with or without assisted embryo selection software) compared to conventional embryo incubation and assessment on clinical outcomes in couples undergoing assisted reproduction, conclude that there is insufficient evidence of differences in live birth, miscarriage, stillbirth or clinical pregnancy to choose between the time-lapse systems and conventional incubation. These results disagree with the meta-analysis of randomized controlled trials done by Pribensky and co-workers[3] showing increasing of pregnancy successful rates, increase of live birth rates and no significant difference in stillbirth. Similar results were found from prospective, blinded, large-sample and multi-center study, carried by Vermilyea *et al.*[37] showing that computer-automated time-lapse analysis correlate well with pregnancy rate and embryo implantation and gave objective and quantitative information to embryologists to improve embryo selection.

The infancy of this fascinating field has, of course, some inconsistence across studies, inviting us to be cautious in the interpretation of time-lapse images and pursue high quality evidence studies in the implementation of this system and based algorithms.

## 3. MACHINE LEARNING

Artificial Intelligence (AI) go backs to 1950´s when computers science pioneers start to hypothesize automate intellectual tasks performed by humans. Thereby, AI include Machine learning (ML) and Deep learning (DL), as such many more other approaches that does not involve a learning process[88]. Refer to Figure 5 with AI, ML and DL timeline and basic supporting idea.

ML is the practice of using algorithms - a set of rules that must be followed to solve a specific problem, to analyse data, learn from that data, and then make prediction about new data[89,90]. The supporting idea of ML is learning from the data and the ability to derive predictive models without a need for strong assumptions about underlying mechanism[91]. The performance and accuracy of the ML algorithms depend on how well trained was to classify and process information. Typical ML workflow involves data harmonization, representation learning, model fitting, validation, deployment and updating. The objective of harmonization step is the transformation of the information in order to extract relevant features (representation learning) for training the model (model fitting). A final step involves testing and evaluation (validation) of the training model[92]. ML algorithms comprises four broad categories: unsupervised, supervised, semi-supervised, and reinforcement learning.

### SUPERVISED MACHINE LEARNING

In this kind of learning process, the goal is to build a concise model of distribution of class labels in terms of predicted features. The resulting classifier is then used to assign class labels to the testing instances where the values of predictor features are known, but the class of the label unknown[93]. It consists of learning to map input data to known targets, given a set of examples. That is, having an input variables (x) and an output variable (y), is used an algorithm to learn the mapping function from the input to the output such as $y = f(x)$.

This type of learning algorithms has a supervisor that assure and supervise the process. The predictions from training data by the algorithms is corrected by supervisor until reaches optimal performance[90,93]. Supervised machine learning refers typical to regression and classification. That is, prediction on continuous scale and prediction on categorical scale. It includes some popular algorithms such as regression (linear and non-linear ), random forest (RF) support vector machines (SVM) and Artificial Neural Networks (ANN) often used in prediction and classification problems[88].

### UNSUPERVISED MACHINE LEARNING

Unsupervised machine learning seek interesting transformation of the input data with no helping targets for data visualization, denoising clustering detection, etc., to better understand the associations presented in the data[88]. The goal is typically improve our insight of a dataset before attempting a supervised learning approach and is usually indicated for data exploration. Examples of this type of learning algorithms are dimension reduction to discover the rules under a large amount of data and clustering to discover inherited groupings in the data[94]. This type of ML includes Apriori Algorithms for association rules learning and K-means for clustering problems[94,95], latent structures projections and principal component analysis.

### SEMI-SUPERVISED MACHINE LEARNING

Semi-supervised learning is a learning paradigm concerned with the study of how computers and natural systems such as humans learn in the presence of both labelled and unlabelled data. Traditionally, learning has been studied either in the unsupervised paradigm (e.g., clustering, outlier detection) where all the data are unlabelled, or in the supervised paradigm (e.g., classification, regression) where all the data are labelled[96]. This type of algorithms works better when using small amounts of labelled data and a large amount of unlabelled data[97].

The goal is to understand how combining labelled and unlabelled data may change the learning behaviour, and design algorithms that take advantage of such a combination. Over the years has aroused great interest in machine learning and data mining because it can use readily available unlabelled data to improve supervised learning tasks when the labelled data are scarce or expensive[96].

### REINFORCEMENT MACHINE LEARNING

Reinforcement learning begun capture the attention after the success of Google DeepMind[98]. In this type of ML, the algorithm learns how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm. In reinforcement learning, an agent receives information about its environment and learns to choose actions that will maximize some reward[99]. Currently, deep learning is enabling reinforcement learning to scale to problems that were previously intractable, such as learning to play video games directly from pixels. Are also

applied to robotics, allowing control policies for robots to be learned directly from camera inputs in the real world[99].

### 3.1. DEEP LEARNING

Deep learning (DL) is a sub-field of machine learning were algorithms or models are based on structure and function of human brain neural networks. Is a mathematical framework for learning representations from data. The networks are called Artificial Neural Network (ANN). ANN is computing systems comprising connected units of nodes. The nodes are organized into layers. If the ANN has more than one hidden layer is called a deep ANN. The signal received by a node is processed and transmitted to downstream nodes within a network. Adding more hidden layers to the network allows a deep architecture to express more complex structures as the hidden layers capture the nonlinear relationships[100].

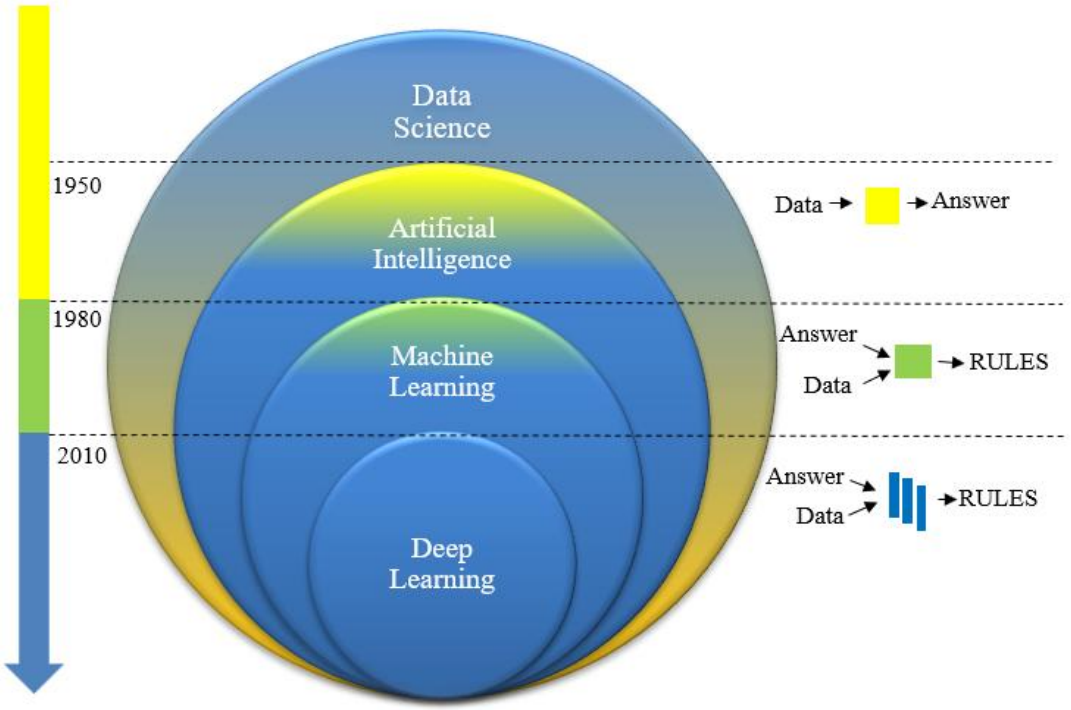



**Figure 5: Artificial Intelligence landscape, showing timeline and differences in data processing.** Deep Learning is a modern branch of machine learning, that learn from representation layers of rules. Artificial Intelligence use data and rules to program algorithms, instead machine learning feed in data and answer to develop algorithm.

DL differs from traditional ML in the number of layers, their connections and how representations are learned from raw data[101]. Indeed, DL allows models with multiple processing layers based in neural networks to learn representations of data with multiple levels

of abstraction[100]. In DL process every layer produces a representation of the observed patterns from data of previously layer by optimizing a local unsupervised criterion. The key aspect of deep learning is that these layers of features are not designed by human engineers, instead they are learned from data using a general-purpose learning procedure[100] That is, DL completely automate the feature engineering, instead manually engineer good layers of representations for the data all features are learnt in one pass rather than having to engineer them, often replacing sophisticated multistage pipelines with a single, simple, end-to-end deep-learning model[88].

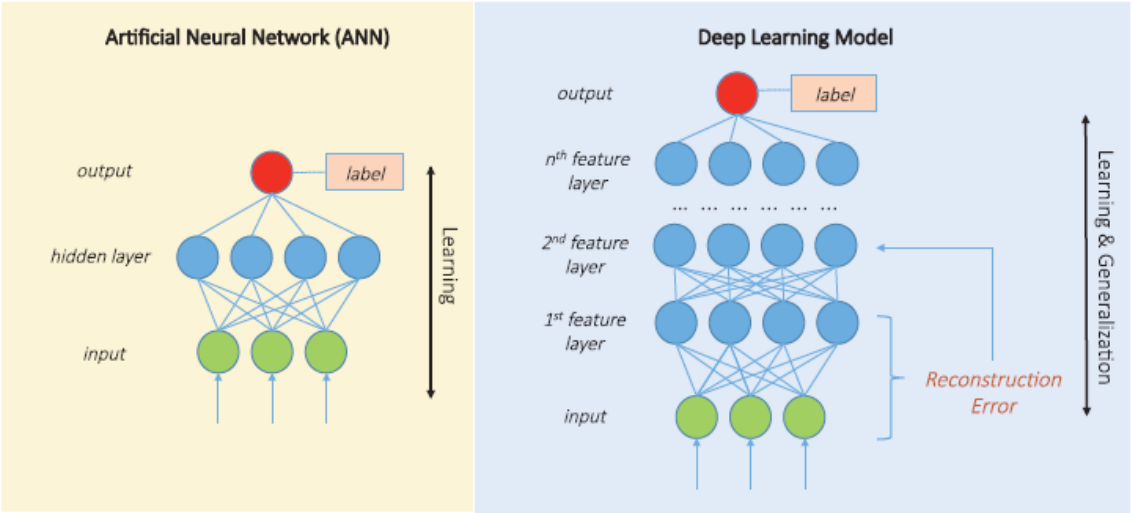



**Figure 6: Architecture of Artificial Neural Network and Deep Learning.** Traditional ML have a basic ANN architecture with three layers of representations toward the final outputs, instead DL model has multiples layers of neural networks (deep in DL). These layers of representations allow to be efficiently tuned and extract deep structures from inputs data to serve as high-level features for a better prediction. Adapted from[101].

As fast-growing branch of machine learning, the representations layers of DL tries to model hierarchical features behind the raw data, being images[102,103], objects[104], sounds[105,106], text[107] or language[108], and classify then by stacking multiple layers of representations.

Google DeepMind AlphaGo project, Google translator, Google´s street view and image search engine, Android voice recognition[109], Microsoft real time language translation or Apple´s virtual assistant Siri, are all example of successfully applications of DL to ruling Big data for competitive advantages[110].

DL has been widely used in medical and clinical imaging to automate and extract relevant features. Example applications include the use of computed tomography images for classification of interstitial lung diseases[111] and for anatomical organ or body-part-specific

classification[112], x-ray images for classification of tuberculosis manifestation[113], colour fundus images for detection of haemorrhages[114] and retinopathy diabetic[115]. Or the use of Magnetic Resonance Imaging for early diagnosis of Alzheimer disease[116], to predict risk of osteoarthritis in knee cartilage[117] and to segment multiple sclerosis lesions[118]. In genomics it was been used to predict the splicing activity of individual exons[119], specificities of DNA and RNA binding proteins[120] or chromatin marks from DNA sequence[121]. This unprecedent success of DL applications come from advances in central processing units (CPU´s) and central graphing units (GPU´s), availability of large amount of data and developments of learnings algorithms[100,122].

### 3.1.1. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network (CNN) was inspired by Hubel and Wiesel's work on the cat's visual cortex. Is the most employed DL architectures[123]. Its architecture can be defined as an interleaved set of feed-forward layers implementing convolutional filters followed by reduction, rectification or pooling layers. Each layer in the network originates a high-level abstract feature [123]. Given the huge number of nodes and parameters to be trained most of the DL architectures are not proper for multidimensional input correlated data such an image. Different from other deep structures, nodes in CNN extract features of small portion of input image[95]. CNN has been designed to better utilize spatial and configuration information by taking 2D or 3D image as input data[122,124].

Standard CNN are composed structurally by convolutional layers interspersed with polling layers, followed by fully connected layers. And, in some cases, by a softmax layer. The convolution filters are applied many times to an image, resulting in series of overlapping receptive fields – the input image is convolved using several small filters, the resulted image is subsampled and are a new input image for the next convolution layer. The process is repeated until top quality features can be extracted. At the end CNN use fully-connected layer to convert the feature format to 1D for final classification (for full review of CNN and DL structures see [95,100]. Typical CNN architecture (Figure 7) involves convolutional, pooling and fully connected layers, Rectified linear unit (ReLU) activation function, loss function and batch normalization.

A **convolution layer** is largely used in computer vision and image analysis. Comprises several small two-dimensional filters and feature detectors. These filters are learnt as part of the network training which is known as representation learning – the training is done alongside

the classifier training. These learnt filters are convolved with the input image and the resulting feature responses are passed upstream to the next processing layer[125]. These convolutional filters share the same parameters in every small portion of the image, reducing the number of hyperparameters in the model[124]. The convolution layers are normally interleaved with pooling layers.

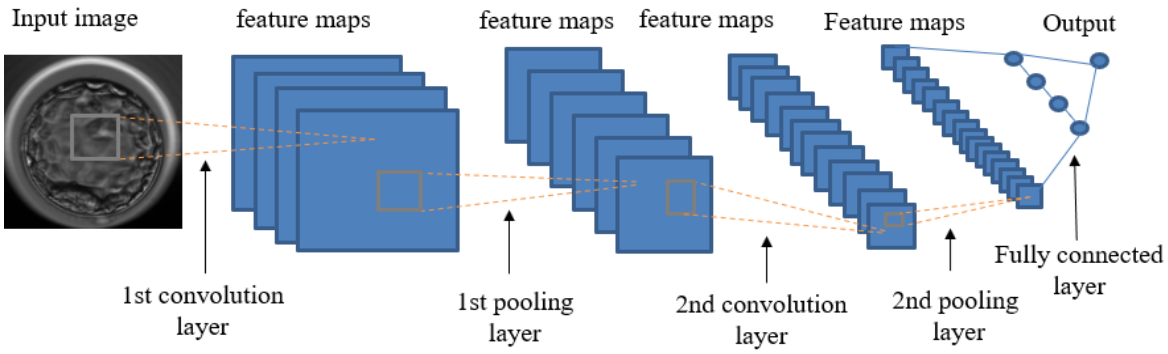



**Figure 7: Typical CNN architecture.** Sequence of convolution and subsample layers of CNN to efficiently process the input image.

**Polling layers**, take advantage of the stationarity properties of images, annotating the mean, the maximum and other statistics of the features at various locations in the feature's maps, reducing the variance and propagating dominant features. The most common polling types are maximum (max) pooling and average or mean pooling. The max pooling capture only the dominant feature response at the poling window. Average polling computes and propagate the mean of all features of the pooling window[122]. Figure 8 shows examples of both. This subsampling also contributes to reduction number of hyperparameters in the model and make the optimization process convenient[100].

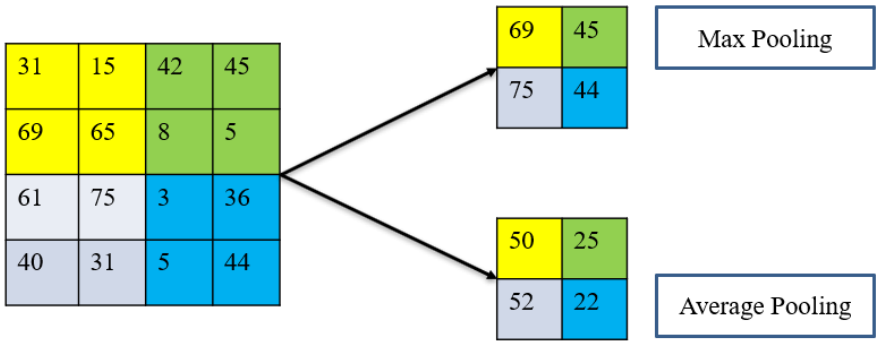



**Figure 8: Examples of simple representation of max pooling and average pooling.**

After interleaved convolution and pooling layers the deep neural network is completed by a **fully connected layer**. As indicated by the name, in fully connected layers all nodes from a layer are linked with elements of preceding layer, resulting in a dense connecting pattern, as shown in Figure 9. To pass from one layer to the next, a set of units compute a weighted sum of their inputs from the previous layer and pass the result through a non-linear function[100]. Nowadays, the most popular non-linear function is Rectified linear unit (ReLU), an activation function.

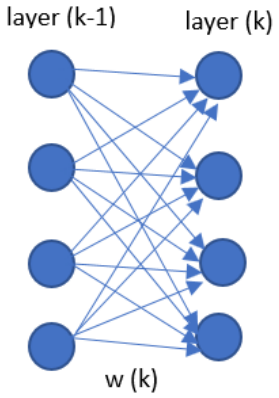



**Figure 9: Graph representation of two fully connected layer (k-1) and (k), connected by a weight matrix, w (k).**

**ReLU** is an activation function that combines non-linearity and rectification layers in CNN. The ReLU formula is $f(x) = max(0, x)$, where $x$ is the input ReLU, as showing in Figure 10[88]. This activation function and its variants shows superior performance compared to hyperbolic or logistic ones. ReLU propagate the gradient efficiently reducing the likelihood of vanishing gradient problem, threshold negative values to zero, solving the cancellation problem and result in more sparse activation volume at its output (providing robustness in small changes in input such as noise) and consist only in simple computational operations been more efficient to implement[126].

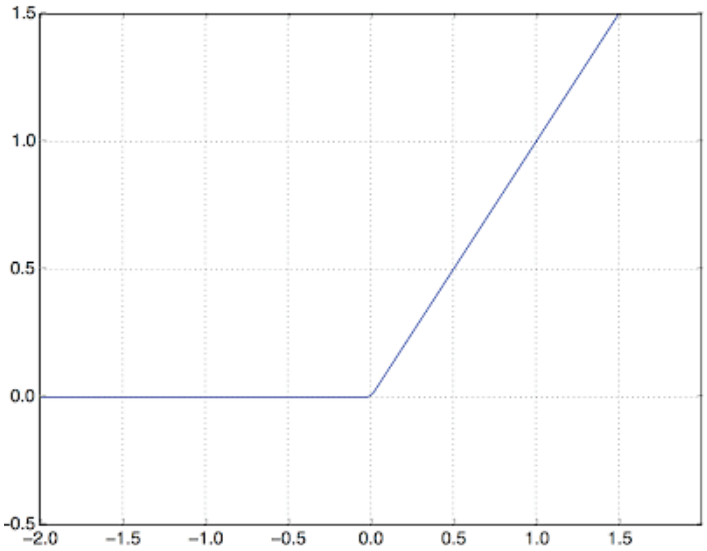



**Figure 10: Visualization of ReLU non-linearity [88].**

Another important component of CNN is **Batch normalization**. Batch normalization normalize or standardize the output distribution of every node in a layer, to achieve stable distribution of activation values throughout training[127]. To do that it normalizes the output of a previous activation layer by subtracting the batch mean and dividing by the standard deviation, putting all data point in the same scale[88]:

i. $z = x - mean/std$, normalize output $x$ from training function
ii. $z * g$, multiple normalize output z by arbitrary parameter g
iii. $z*g + b$, add arbitrary parameter b to resulting product $(z*b)$

Additionally, there are two key concepts to configuring learning process in deep CNN: Loss function and optimizers. The **loss function** is the quantity to minimize during training, thereby represent a measure of success for the task that we trying to solve. Loss function measure the discrepancy between the output of the network depend on the model parameters and the expected results, that is, the true class label in classification tasks, or true level in prediction class[128]. The **optimizer**, in other hand, *specifies* the exact way in which the gradient of the loss will be used to update parameters. The loss function compares the predictions to the targets, producing a loss value that measure the match of the network's predictions and expectations. The optimizer uses this loss value to determine how learning proceeds[88].

### 3.1.2. DEEP LEARNING WITH KERAS AND TENSORFLOW

**Keras** is a high-level neural network Application Programming Interface (API), written in Python that can run on top of TensorFlow (among others, like Theano). This modular, extensible and open source API enables fast execution with deep neural networks and allows easy and fast prototyping. Keras offers consistent and simple API, minimizes the number of user actions required for some common tasks[129]. Provide high-level building blocks for developing deep-learning models. Require a specialized well-optimized tensor library do that, serving as the backend engine. One of this backend deep-learning execution engines of Keras is TensorFlow[130,131]. Over TensorFlow, Keras can run on both CPUs and GPUs. However, on GPU TensorFlow encompasses a library of well-optimized deep-learning operation[132].

**TensorFlow** is an open-source software library, developed and released in 2015 by Google for systems capable of building and training neural networks, using different datasets. Initially, the main goal was to detect and decipher patterns and correlations, similarly to the learning and reasoning of humans[133]. TensorFlow uses a unified dataflow graph to represent both the computation in an algorithm and the state on which the algorithm operates[132]. It uses numerical computation and data flow graphs[134].

TensorFlow name derived from the operations that neural networks perform on "tensors" that are multidimensional data arrays. Since is designed for numerical computation it uses nodes and graph edges. The nodes represent mathematical operations and the graph edges represent the tensors communicated between them[131,134]. Also, this framework improves efficiency and modularisation in distributed computation by TensorBoard, a supporting tool for in-depth visualization of training process, facilitating global representations of complex model, debugging and checking along development of the model[135] (Figure 11). To perform an action in TensorFlow, is required to perform some other tasks before the execution of the action, and in Keras these tasks can be performed with a simple line of code[129,130].

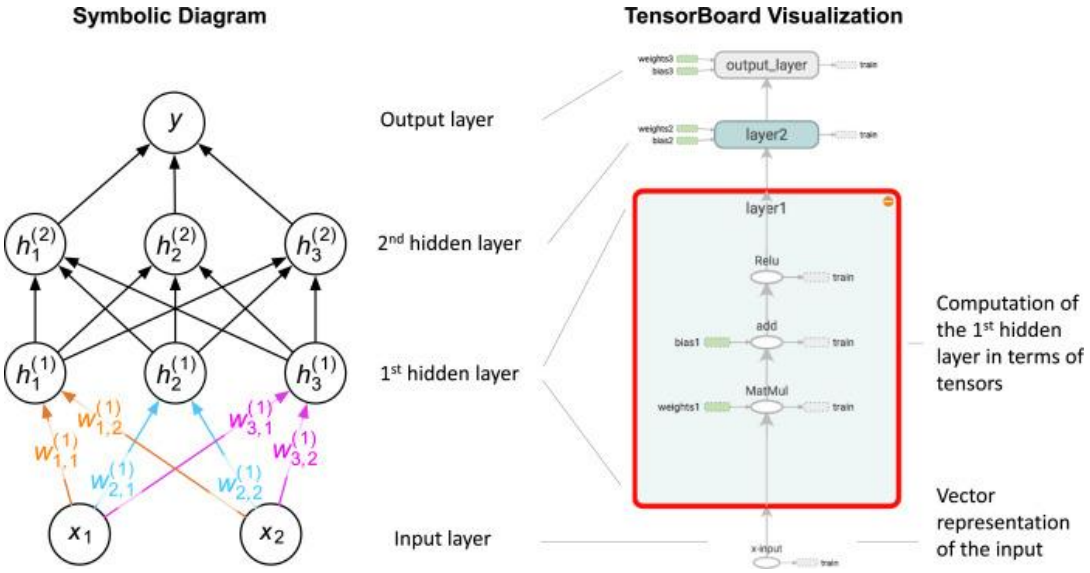



**Figure 11: Deep neural network with two hidden layers in TensorBoard.** TensorBoard provide supporting tool for deep visualization of training process and simplify the model representations. Adapted from [135].

## 4. AIMS

The objectives of this thesis project were to shows the viability of deep learning approach for embryos selection, construct an analysis pipeline for embryos classification and acquire acquittance in vast and complex domain such as ML. The framework is based on TensorFlow open source and Keras libraries and are developed to extract and classify features associated with embryos outcomes to assist the clinician in selection of embryos with better prognosis.

# II. MATERIALS AND METHODS

## 1. IMAGE ACQUISITION AND PREPARATION

This study includes 117 Human embryo images, gently provided by Ferticentro, stored in JPEG format and labelled by Embryoscope™ software classified from 1(worst quality) to 6 (best quality)(Figure 12A). The images were captured using EmbryoScope time-lapse system (Vitrolife™, Sweden), with a built-in microscope. The system captures the images 110 h post-insemination, using single red LED (635 nm) every 20 minutes, with seven focal depths of the embryo taken each time.

There were 79 day-5 embryos images scored between 1 and 4, and 38 images scored as 5 or 6 (Figure 12B). Two images were removed due to dark background. We labelled the images classified by the software score 5 and 6 as prone to be implanted and categorized as good quality and 1- 4 as not prone to be implanted as poor-quality ones, following the Adolfsson *et al.* study in which highest live birth and pregnancy rates was superior with KIDScore 5 and 6 embryos, decreasing with KID 4 and 3 and being KID 2 and 1 similar live birth and pregnancy rates[136]. The KID scores reflect implantation potential by analysing large database of embryo development with known clinical outcome. The models are developed by analysing how embryo morphokinetics, cleavage patterns and morphology correlates with implantation outcome after transfer[87].

For this work all images are directly used as the input data, without any pre-processing methods, such as deconvolution[137] or algorithms to identify cells manually[138]. Images modifications are warranted to improve training process by artificially increasing the search space with image variations. The images are split into training, validation and test groups. 80% are allocated to training dataset and the remaining 20% to the test dataset. The datasets did not overlap. The framework is shown in Figure 13.

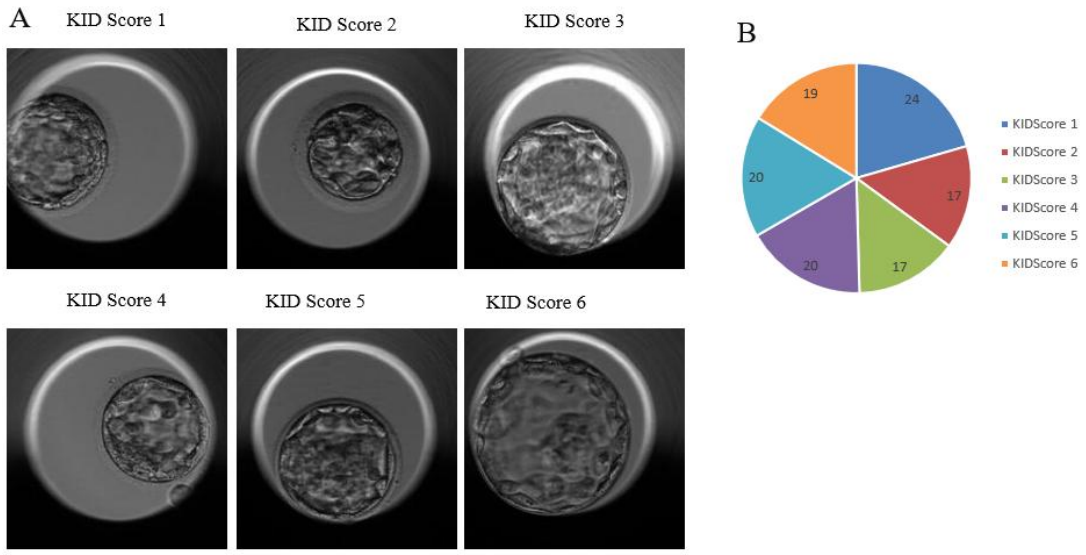



**Figure 12: Pattern of collected images.** A. Images representing each KIDScore group. KIDScore 1 embryos are the ones with too fast start up to three cells. The group 2 embryos have too slow initial development. KIDScore 3 embryos have irregular divisions with increasing development speed between the two and five-cell stages. KIDScore 4 embryos are composed of two types of embryos: those that have irregular divisions and those that do not reach eight cells prior to 66 hours post insemination. Group 5 and 6 are those that have passed all avoidance criteria[136]. B. Images groups distributions. Images 1, 2, 3 and 4 are classified as poor quality ones, that is not prone to be implanted. Images 5 and 6 are classified as good quality ones and prone to be implanted. Images kindly given by Ferticentro, S. A.

## 2. MODELS DEFINITION AND TRAINING

To implement our framework, we used TensorFlow version 1.1.0 (Abadi, 2015) and the Keras 2.2.4 for defining, training, and evaluating models. Training of our CNN method was performed on a server running the R version 3.5.1 (2018-07-02) under Windows 10 Home, 64-bit operating system. This server is powered by four NVIDIA GeForce GT 620M with 8 GB of memory and 1.70-GHz Intel® Core ™ i5 CPUs.

Our Deep Learning framework is based on CNN, as our input data are raw image instead of extracted features, which reduce the dimensionality of learnable parameters, and alleviate the training process, through several constraints on the synaptic weights[139]. We resize the input image to 28 x 28, exploiting local features with convolutional receptive field of 3 x 3, reducing parameters and to force network going deeper. Resizing the images data make them more analytically computational, less expensive and thus faster, gaining space storage, transmission time and management efficiency and querying[124].

The CNN architecture comprise several convolutions to pass the results to the next layer[125], pooling layers to combine the outputs of nodes into single nodes[100,122] and fully connected layers that is the outputs. The first two 2D convolutional layers are combined with a max pooling operation, with the pool size 2×2 to reduce the size of feature maps down to two times[88], two ReLU activation functions at each convolutional layer, as neuron model to avoid vanishing the gradient problem and a dropout layer – a regularization method based on efficient ensemble learning, with dropout rate of 0.25 to prevent overfitting[125]. The 2D convolutional layers will deal with the 2-dimensional matrices input images. The model follows with the repetition of the previously architecture. The model architecture ends with a fully connected layer and a softmax output layer to map the nonnormalized output to a probability distribution over predicted output classes, which means that the model will make its prediction based on the option with highest probability[95,100]. The final layer outputs a length 2 vector (probabilities for each class 0 and 1) using a softmax activation function (Figure 14, model architecture).

To compile the model three parameters were used. The optimizer to determine how fast the optimal weights are calculated, loss functions and accuracy metrics to accuracy score when we train the model. As optimizer we chose `Adam` which adjust learning rate throughout the training. The method is straightforward to implement, is computationally efficient, has little memory requirements, are appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients[140]. The hyper-parameters have intuitive interpretations and typically require little tuning[140]. To estimate the error of the model and update the weight to reduce the loss of next evaluation we choose binary cross entropy classification, that is the state-of-art for binary classification problems.

The `fit()` function is used to run CNN on our training data. The object is our model, and x and y coordinates are our training data in a list of inputs images. The `batch_size` reflects the number samples per gradient update within each epoch. `Epochs` are used to control the number training cycles. Typically, we want to keep the batch size high since this decreases the error within each training cycle (epoch). We also want epochs to be large, which is important in visualizing the training history. The `validation_split` is slated to 0.20 to include 20% of the data for model validation, which mitigates overfitting. The training process runs in a few minutes.

Due to a low number of images and to increase artificially the number of image data, to reduce overfitting, as well to improve performance of machine learning framework in imbalanced class problems, we used “data augmentation”, that enlarge the training dataset.

Data augmentation constitute also a good technique to make our model invariant to changes in size, translation, viewpoint or illumination[141]. The generation of new labelled images can be done rotating the original images, flipping horizontally or vertically the images, zooming, width or height shift range the images, creating a warped version of training data, using `image_data_generator` from Keras. In our data augmentation the transformed images are generated in R code on the CPU, with very little computation and no need to be store on disk. Prior to augmented training process, we additionally use manually transformed image on test set to verify the adverse examples impact in overall training process and study the effect of balanced dataset on designed neural network.

## 3. MODEL TESTING AND EVALUATION

The performance of the model was not the aim of this study (which was constrained by the low number of images), but instead to demonstrate that deep learning could be applied to help to select the best embryo in clinical practice. Therefore, the assessment gives us the state and response of framework to different training parameters. To evaluate the performance of proposed framework is required a collection of statistics parameters derived from confusion matrix – a specific table that allows the visualization of the performance of an algorithm (Table 4). We use accuracy measurement to identify the portion of image correctly identified and AUC (ROC Area under curve), to measure the trade-offs between sensitivity and specificity in our binary classifier. Accuracy is a critical measure for evaluating the performance of a classification algorithm. When all instances in a dataset have the same weight, the accuracy of a classifier on a dataset is defined as the number of instances predicted correctly over the total number of instances[142]. The receiver operating characteristics (ROCs) reflect the plotting of true positive rate (TPR, specificity) against false positive rate (FPR, or 1 - sensitivity) at various threshold settings[142,143]. The metrics are used to measure classification performance and accuracy of classification model. The accuracy is measured by the area under the ROC curve (AUC)[143].

$$Accuracy = \frac{Number\,of\,correct\,preditions}{Total\,number\,of\,preditions} \quad (142)$$

For binary classification, accuracy can also be calculated in terms of positives and negatives, such as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP stand for True Positives, TN for true negatives, FP and FN for false positives and false negatives, respectively. The confusion matrix summarizes the binary classification results, where FP and FN correspond to the classification errors.

**Table 4:** Confusion matrix

| | **Actual class 0** | **Actual class 1** |
|---|---|---|
| Predicted class 0 | True negative (TN) | False negative (FN) |
| Predicted class 1 | False positive (FP) | True positive (TP) |

Due to limited image dataset and to obtain a fairly reliable accuracy estimation, as well as to ensure that every embryos image from the dataset pass to the training and validation sets, thus reducing the classification error, we employ a k-fold cross validation. In this technique the model is trained a total of k times, leaving each time a fraction of 1/k for validation, leading to k distinct folds. Each fold provides the model performance metrics summary, and the overall assessment of the model performance are the average of each fold result[142]. Similarly, since labelled class 0 is twice the class 1 embryos images, we pay attention to the effect of class imbalance on test dataset in each fold accuracies.

For a given training conditions we present a learning curve. The learning curve depicts the plot of training and validation accuracy as function of the number of training data[91]. It measures how well the model perform with the training and unseen data, comparing training and testing accuracies and post the degree of convergence between them. The train learning curve are calculated from training dataset and shows how well the model is learning. The validation learning come from the hold-out validation dataset and represent the generalization process of the model[91,94].

# III. RESULTS AND DISCUSSION

## 1. THE FRAMEWORK AND DESIGNED DEEP LEARNING NETWORK

The overall framework of our method is shown in Figure 13. An image obtained from Embryoscope$^{TM}$ time-lapse system and ranked from 1 to 6 by KIDScore software, are labelled as poor quality (0) and good quality (1), for embryos ranked from 1 to 4 and from 5 to 6, respectively, based on their live birth rate and pregnancy likehood. The labelled image, with no further transformation but the ones required to improve the network performance, is presented to designed network that generate a numeric values representing the morphology of the embryos and provide an output class that represent the current quality of the embryos, that is the likehood of an embryo to be implanted and results in a pregnancy and successful live birth. In our study, the embryo images are raw time lapse images taken after fertilization with no further segmentation by advanced image analysis techniques, thus the images quality could have more impact on the embryo grading outcome. The training and validation phase are done under the designed network shown in Figure 14, at which the learnable parameters are adjust as described in method section 2.

The model architecture is shown in Figure 14, as well as tensor board graph in Figure 15. The designed deep convolutional network comprehends a combination of convolved layers with varying output channels and kernel sizes, polling layers, rectified linear unit layers, dropout layers, flattened layer, fully connected layers and a softmax layer that shows the probability of class from an embryo image. Each input is propagated through separate convolution, pooling and dropout layers. The output of each layer is combined to form a flatten vector, reaching a fully connected layer and dropout layers. A new fully connected dense layer and softmax function output the class probability.

From learnt a feature in a specific location of the image by convolution layer, adding a fully connected layer allow to recognize these features anywhere in each image. Increasing the efficiency in embryos images processing and allow the model to generalize better in a few sets of images samples. By duplicating the convolution layers, the first layer will learn small local features and the second layer will learn larger patterns of first convolved layer, allow the CNN to learn spatial hierarchies of patterns[100,128]. The pooling layer insertion after two convolutions allow the parameters reduction in the model, by down sample the data representation, reduce the computational cost in the network[88] and control the overfitting[144]. Dropout, in turn, will deactivating a percentage of weights of the network (randomly) during training stage by adjust

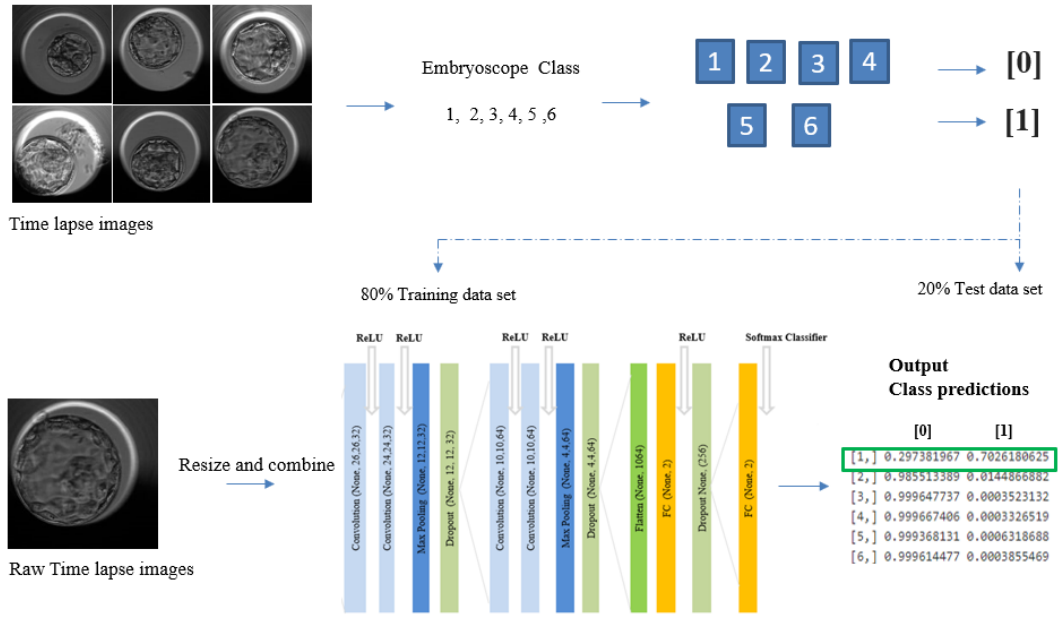



**Figure 13: The proposed framework.** This flowchart illustrates the design and assessment of proposed framework. Human embryo images are provided from the embryology lab and classified by Embryoscope time lapse system software KIDScore. Then classified embryos images are labelled as poor quality [0] or good quality [1], based on their live birth and pregnancy rates. The labelled class are presented to designed network to generate the numeric values that represent the morphology of the embryo. The network evaluates the values and provide an output classes that represent the quality of the embryo.

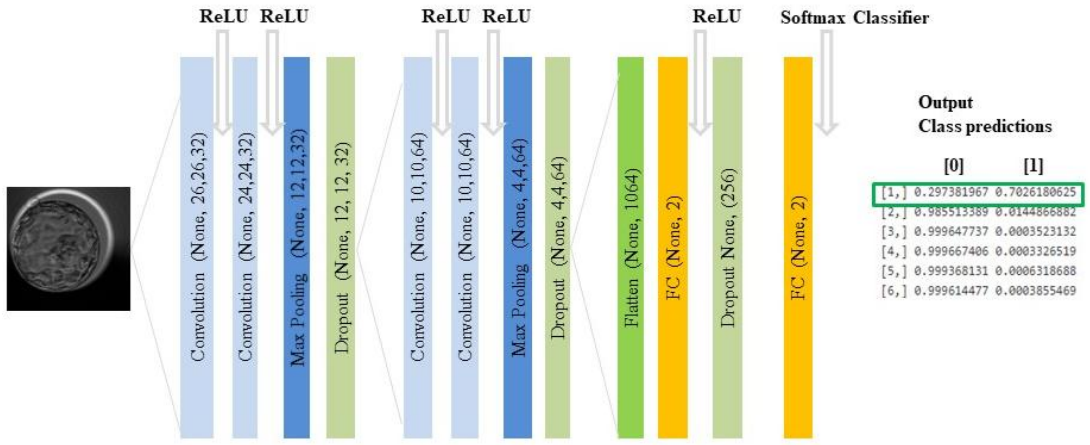



**Figure 14: Simplified designed architecture.** The day-5 embryos are first processed by two rounds of two convolutions followed by max polling, dropout regularization and ReLU layers. The (4, 4, 64) outputs are flattened to (1064) before through dense connected layer, both under dropout regularization and ReLu non-linearity, follows. The last hidden dense layer, under a softmax function, output the class scores. These scores are used to calculate the loss function and to make class predictions in the training and test stages. FC stands for fully connected layer.

their values to zero[144]. The network receives as input raw images of size 28 x 28 with 3 channels, each for a color Red, Green and Blue (RGB) and a dropout with 25% of keeping probability during the training phase. ReLU was used as activation function for all layers except the output, in which was used softmax function (Figure 14). The full description of the network architecture is in method section 3.

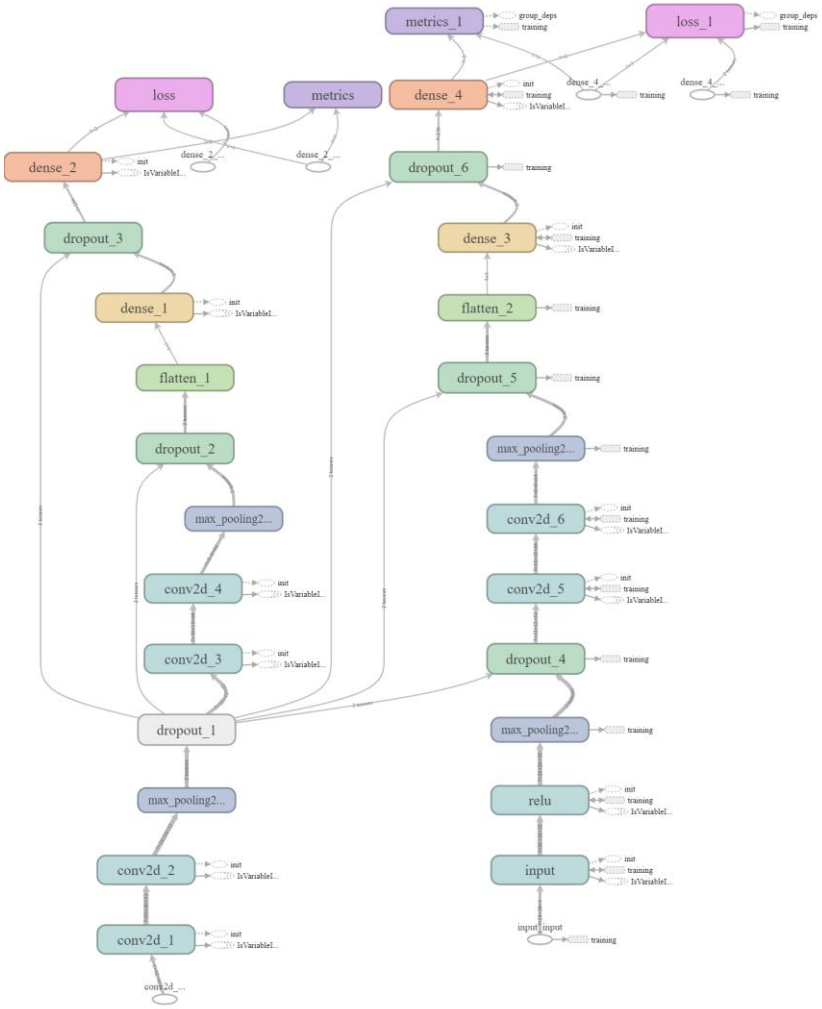



**Figure 15: TensorBoard visualization of training process and model architecture.** The TensorFlow Graph shows the presented convolutional neural network for embryos images classification, displaying a dataflow between groups of operations, with auxiliary nodes extracted to the side.

## 2. MODEL TRAINING AND EVALUATION

After defining and performing a fine-tuning for all the layers of the network parameters, such as image size (100 x 100, 40 x 40, 28 x 28 pixels), batch size, dropout regularization (0.0

to 0.5) to find proper CNN structure, we trained our deep learning network to classifies the training and test sets images. The results show that the trained framework was able to identify good and poor-quality embryo with up 99% of accuracy for training dataset and 86% for test dataset, as show in confusion matrix tables in Figure 16B and 16C, with AUC of 0.52. Therefore, the difference between the accuracy of a model on examples that it was built on and examples that it have not seen before, coupled with low AUC value, suggest that model learned rules specifically for training dataset and those rules do not generalize well beyond the train set. This tendency is shown on Figure 16A. The training loss decrease, and training accuracy increase with every epoch. Otherwise, the performance on new data started to stall at epoch 40, compared to continue improvement of training data. The model appears to underestimate the good quality class of embryos . Should be the case that learned specific rules on train dataset work against the test dataset.

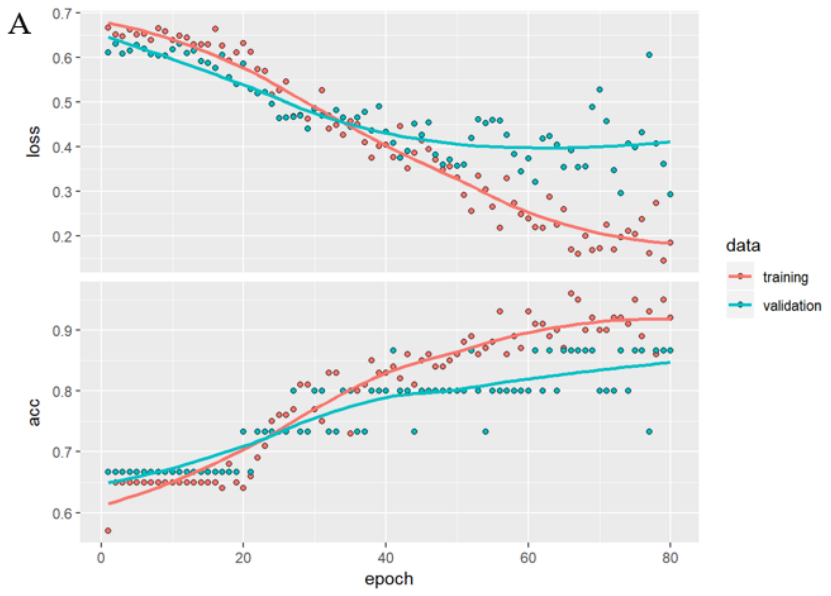



**Figure 16: Training and validation metrics of the network.** Results reflecting the general relations between the input and output data (A). The network was training for 80 epochs with batch size 32 and dropout of 0.25. It fairly predicts the class labels of training images (B) and new images as shown on confusion matrix table for test (C) and the first five training images predictions (D).

The observed overfitting phenomena (Figure 16A) could be explained at large scale due very few numbers of dataset or by complexity of the model. The model excessively adjusts to the training data, that is, consider patterns that are specific to the training data but do not exist or irrelevant on the new data. Consequently, it performs fairly but not full accurately on new data. As stated by numbers of works[111,113,117,118,125], the fewer samples for training, the more models can fit or adjust to the data. When the data increase, fewer model would be able to explain them. Nevertheless, the CNN demonstrated robustness in the consistency of the classification, through several learning sessions.

### 2.1. ADVERSE SAMPLING

Popular techniques to avoid overfitting and improve model generalization include increase training data, regularization such as dropout, reduce network size or add weight regularizer and balance dataset[88]. Dropout and weight regularizer are already include in the network architecture and layers reduction provide no evidence of improvement (data not shown). Next, we try the effect of manually transformed and artificial augmented images on the CNN.

To evaluate the performance of trained network with adverse samples, we add new manually transformed images to the training and test set. The manually transformed images involve single spatial transformation per image, such as rotation, flipping or brighten the images (see Figure 18, for transformed images examples). The predicted class are compared with the true classes to estimate identification accuracy. Our network was capable of predicts the class label of new raw images and can recognize the transformed ones (Figure 17).

Manually transformed images has little effect on CNN performance. The neural network performs almost in the same way as no transformed image dataset. In fact, the major impact is seen on validation performance (17C and 17D) and test data set. When these are coupled to test data there is a loss increasing and a concomitant accuracy decreasing (Table 5). We hypothesize that these transformations are so simple that combined features do not add additional information and/or don´t provide new images that network has not seen before.

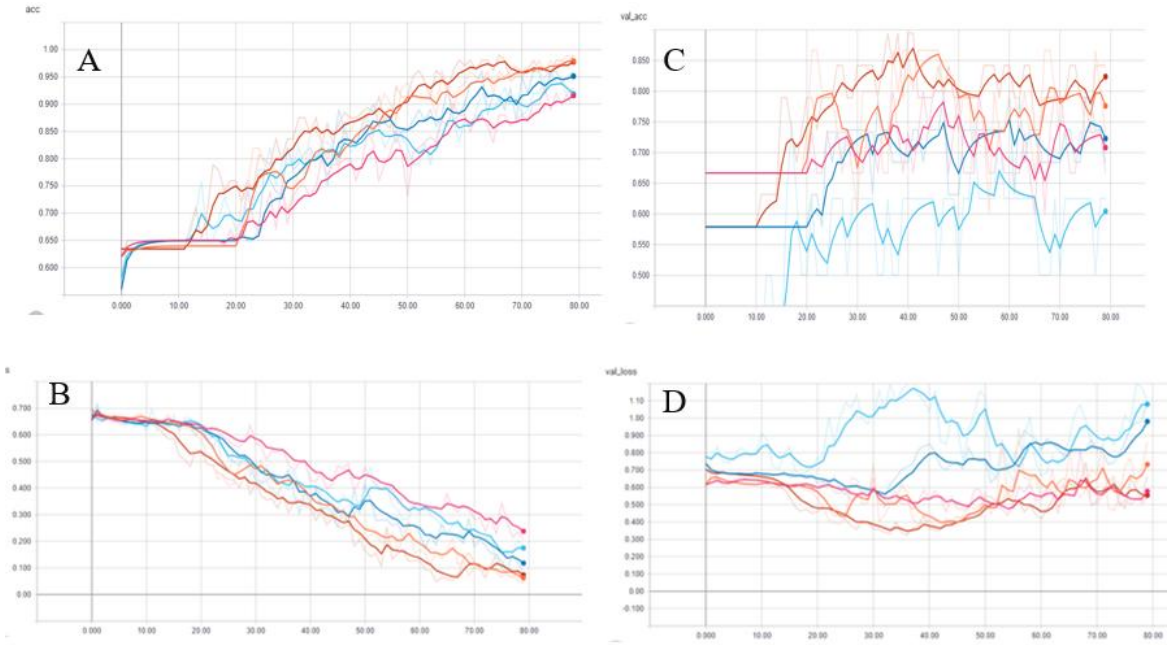



**Figure 17: Tensor board graphs of training network with adverse sampling result reflects the general relations between the input and output data.** The images reflect the overall accuracy (A), loss (B), validation accuracy (C) validation loss (D). The numeric values of train and validation accuracy and loss are shown on Table 5. Tensor board images with runs normal images data (rose), training set with 10% manually transformed images (orange), test set with 25% manually transformed images (blue), Training and test both with 10 and 25% manually transformed images, respectively (red) and test set with 50% manually transformed images (sky blue). It accurately predicts the class labels of new images.

**Table 5**: Loss and accuracy of train and test sets with different dataset type.

| | **Train** | | **Test** | |
|---|---|---|---|---|
| **Dataset type** | Loss | Accuracy | Loss | Accuracy |
| No transformed images | 0.02 | 0.99 | 0.45 | 0.86 |
| Transformed images on training set | 0.02 | 0.99 | 0.86 | 0.73 |
| Transformed images on test set | 0.04 | 0.98 | 1.09 | 0.68 |
| Transformed images on both | 0.02 | 0.99 | 0.05 | 0.84 |
| Test set with half of transformed images | 0.17 | 0.93 | 1.08 | 0.82 |

### 2.2. DATA AUGMENTATION

The network fit well in training data, but the real challenge is generalizing to new data. In deep learning approach the number of images is known to be an important factor to improve accuracy and avoid overfitting and reducing generalization error by simulating realistic

variation of the training data[128]. These artificial variations of images mimic the appearance of future test samples that deviate from the training group.

To address the data scarcity problem, we use artificial images augmentation from Keras, using training images datasets. The test set remain unchanged. In the augmented training process (Figure 19 and 20), every iteration presents a new and random modified version of existing images to the network. So, the model would be exposed to huge amount of possible aspect of the image data distribution and generalize better. We present to the CNN 45% rotated, 20% flipped, random translated images vertically and horizontally, random sheared and zoomed images, as well standardize pixel values across the entire dataset by feature standardization setting "featurewise center" and "feature_std_normalization" arguments on Image data generator. Example of augmented image are shown in Figure 18.

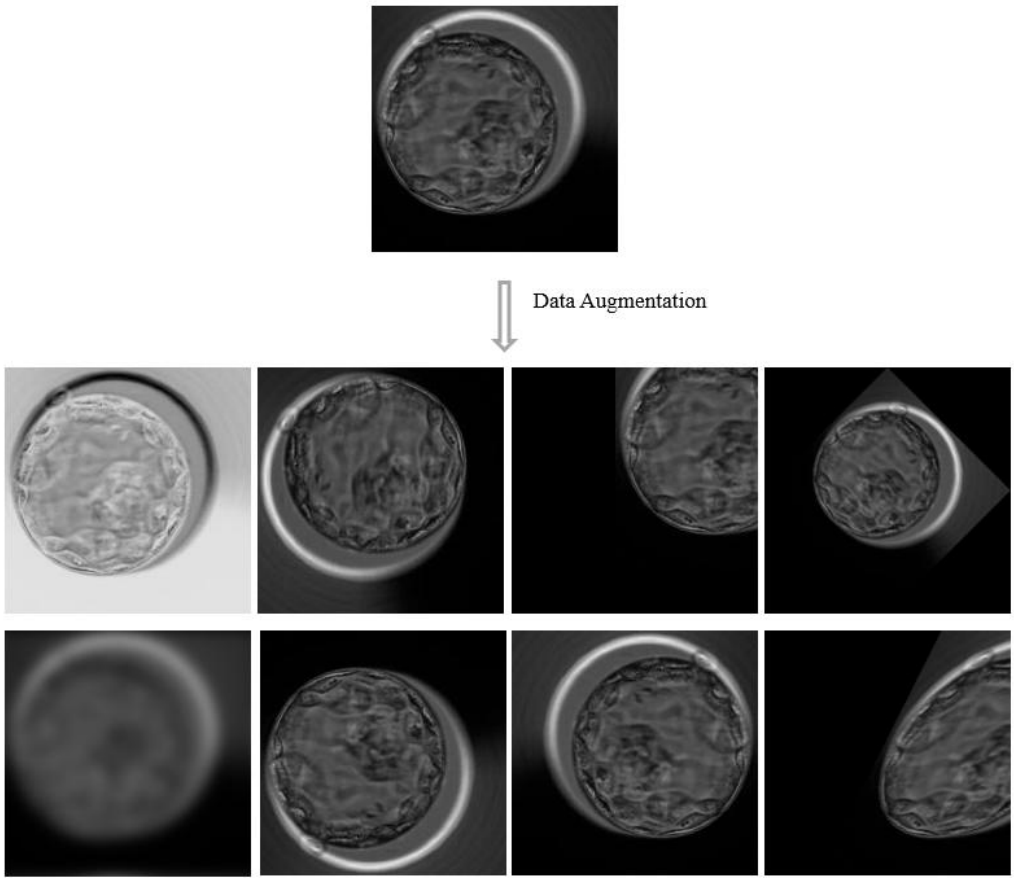



**Figure 18: Data augmentation. Example of data-augmented images.** All the presented images were obtained from a single image by randomly rotating, translating, zooming in/out and horizontally flipping, flopping, sheer or applying filter to the image.

Image data augmentation does not improve the performance of the CNN. It does not increase the network ability to generalize to unseen image variations. In fact, it appears to worsen the overfitting limitation and decrease the overall accuracy. The validation accuracy ultimately reached approximately 35% (Figure 19). It appears that heavily transformations performed by image data generator on Keras, such as shear and zoom range or vertical and horizontal translations, deprive useful information to the model and could act as distracting noise, increasing overfitting and generalization error. Indeed, when these arguments are turned false or eliminated are denoted a network improvement, reaching an accuracy of 0.66 and 0.71 for training and test, respectively. Nevertheless, the training and validation loss remain large (Figure 20). In time-lapse embryos images, orientation, intensity and lateral/horizontal asymmetry sound to be important to the embryo morphology, so heavily augmentation technique might damage the image semantic content and difficult features extraction by the network[145]. It is noted that, at some instances, the augmentation techniques adopted have a large effect in class discrimination and network generalization performance, putting the transformations outside of the range and may not preserve the class specificity.

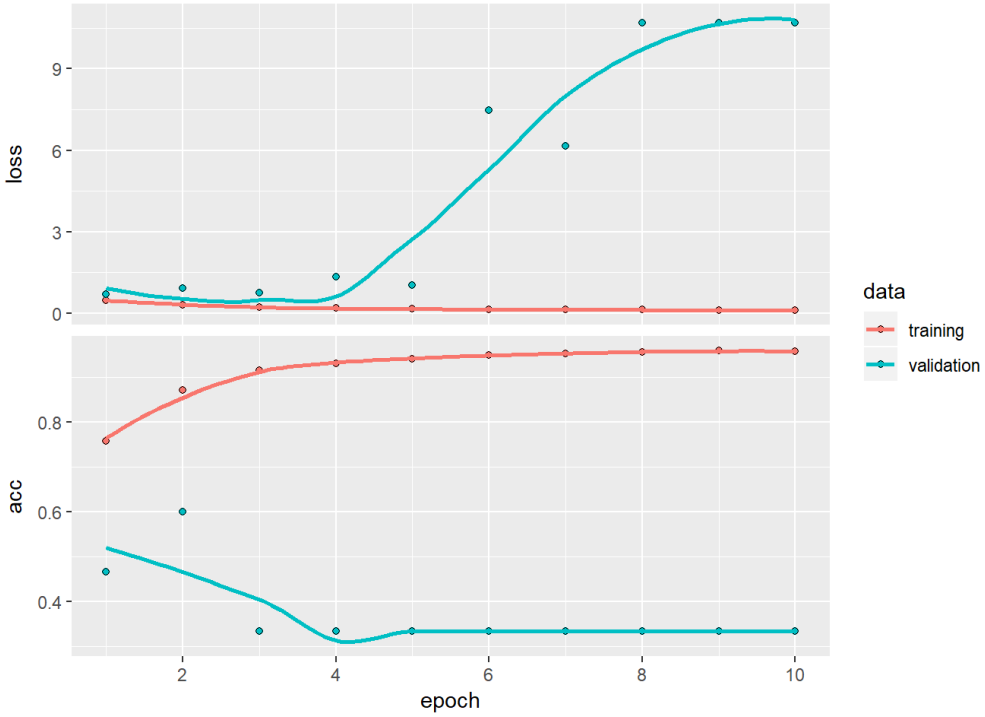



**Figure 19: Training process with artificial image augmentation.** The network was training for 10 epochs with batch size 32, dropout of 0,25, 5000 steps for epochs and validation steps. The overall accuracy was 0,35 for training and test sets.

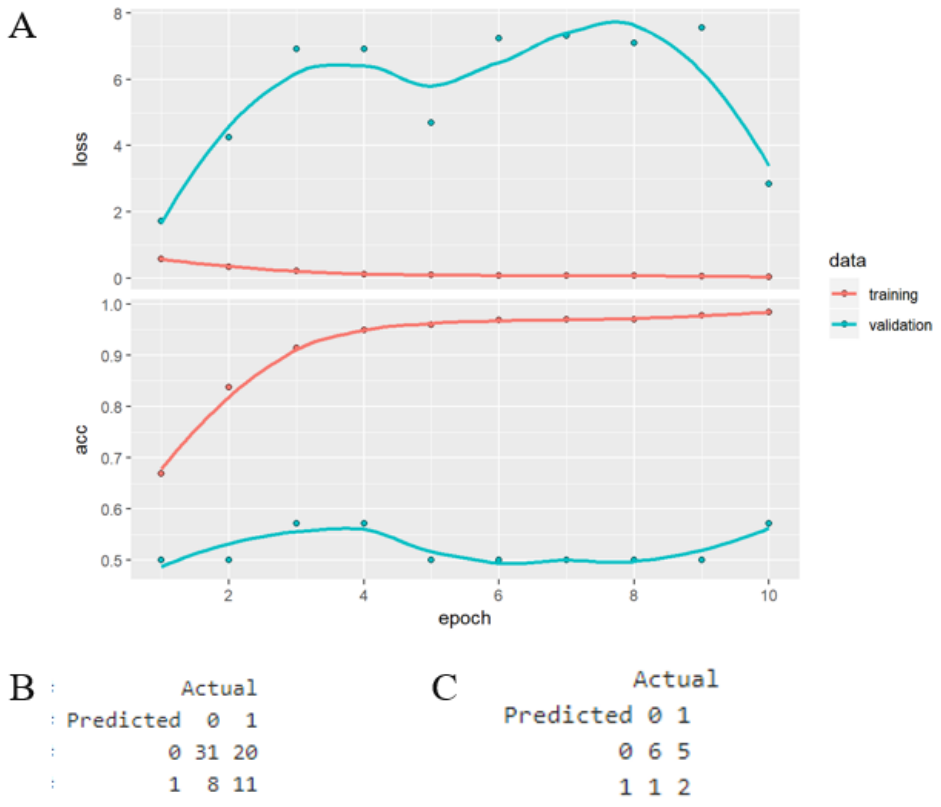
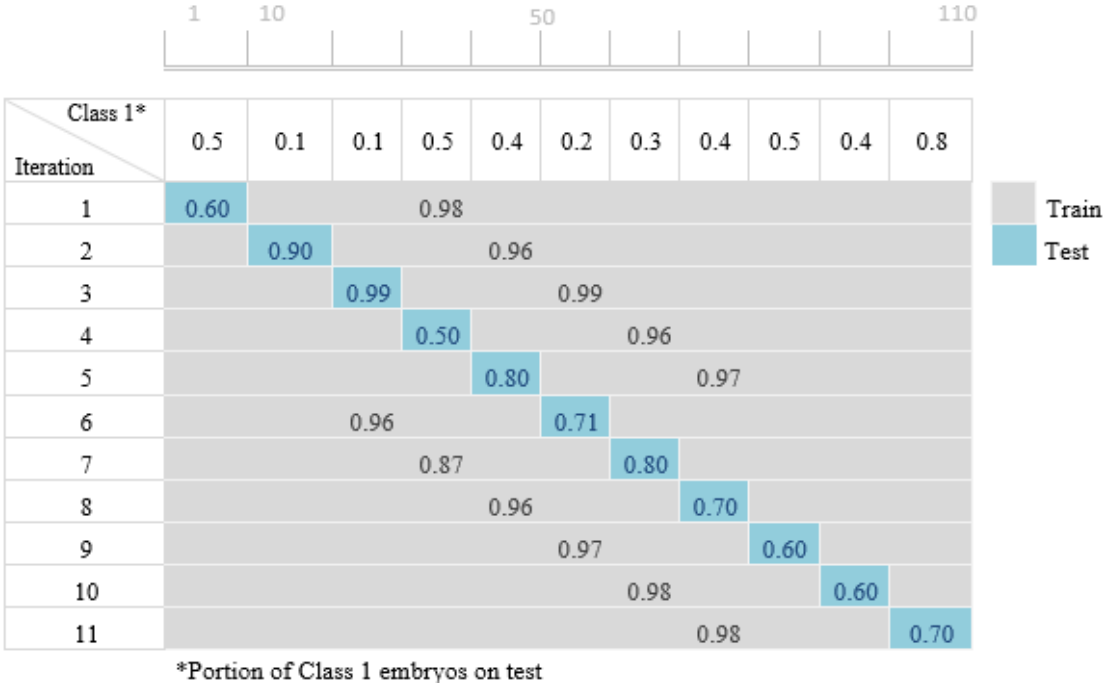



**Figure 20: Training process with artificial image augmentation (balanced dataset).** Heavily augmentation turned off or eliminated. The network was training for 10 epochs with batch size 32, dropout of 0,25, 1000 steps for epochs and validation steps. The overall accuracy was 0.71 and 0.66 for training and test, respectively. It underestimates class 1 embryos as shown on confusion matrix table for training (B) and test (C).

### 2.3. K-FOLD CROSS VALIDATION

Finally, we evaluate the generalization capacity of the model passing every embryos image to training and test sets. The results for each fold are presented in Table 6. The cross validation average accuracies are 0.96 for training datasets and 0.71 for test. As noted previously, the model appeared to underestimate the class 1 embryos. A test accuracy reaches 80% with balanced number of good and poor class embryos and increase to 99% when presented larger group of poor quality embryos on test. It seen that the model does not learn all the features that characterize the embryos prone to be implanted, due to few numbers of these class of embryos presented during training process. In other words, for the presented input images datasets, the imbalance of class should cause an over-classification of poor quality class embryos due its prior probability[146] and underestimate good quality class embryos.

**Table 6:** Accuracy for each iteration of cross validation study




A comparison of all training process shows some consistent results (Table 6). The model performs well on training dataset; therefore, the model is overfitting, mostly due to reduced number of training dataset. Exception to augmented training process where are show and underfitting phenomena (Figure 19 and 20). The model seems do not recognize the morphological features on the random transformed images presented. On other hand, the training data seem not to be enough to learn the required embryo features.

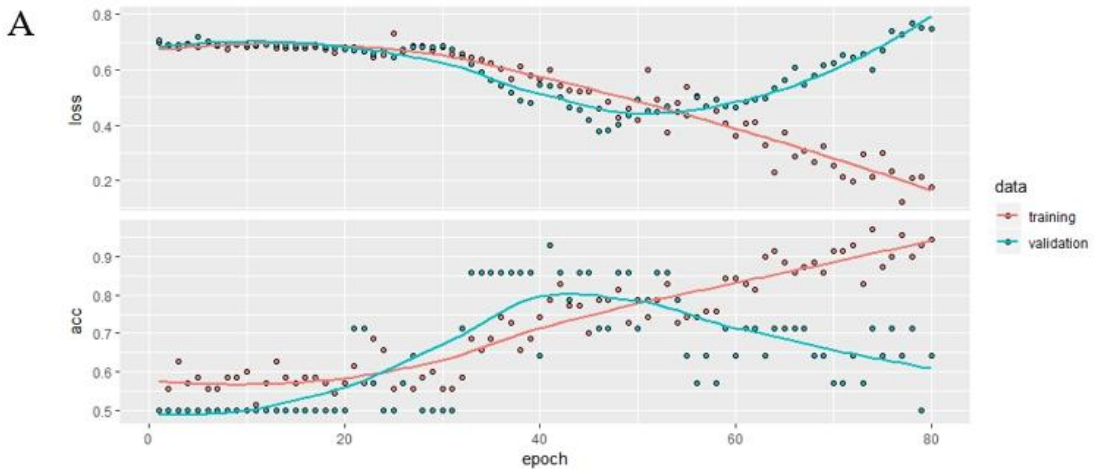
As shown in Table 7, we try to balance the dataset even considering that we are reducing the overall training dataset. And, as expected doesn´t improve the generalization properties but clearly reduce the good quality class underestimation (Figure 21).

**Table 7:** Summary of training and test accuracies*

| | **Training Accuracy** | **Test Accuracy** |
|---|---|---|
| Normal data | 0.99 | 0.86 |
| Data with adverse sampes | 0.99 | 0.84 |
| Augmented data | 0.71 | 0.66 |
| Cross validation | 0.96 | 0.71 |
| Balanced data | 1 | 0.71 |

* The best results are shown, otherwise for Cross Validation the mean of all the iteration.

To observe the effect of class balance, we try different the loss of function and balance the dataset approximating the classes numbers data[88,146]. Change to new loss function can allow the minority samples to contribute more to the loss[146]. To training balanced dataset we reduce the total of training datasets to 70. From all the studied panel of loss of function binary cross-entropy perform better (Figure 21). The accuracy decreases to stable 71%, compared to imbalanced data (Figure 16 and Table 7), for test data set and accurately predicts all training data (Figure 21).




B

| Predicted \ Actual | 0 | 1 |
|---|---|---|
| 0 | 39 | 0 |
| 1 | 0 | 31 |

C

| | | Predicted_class | Actual |
|---|---|---|---|
| [1,] | 0.0057438952 0.9942560792 | 1 | 1 |
| [2,] | 0.9983995557 0.0016004080 | 0 | 0 |
| [3,] | 0.7885764837 0.2114234865 | 0 | 0 |
| [4,] | 0.8291667700 0.1708332598 | 0 | 0 |
| [5,] | 0.9872165322 0.0127834640 | 0 | 0 |

D

| Predicted \ Actual | 0 | 1 |
|---|---|---|
| 0 | 4 | 1 |
| 1 | 3 | 6 |

E

| | | Predicted_class | Actual |
|---|---|---|---|
| [8,] | 0.0766323134 0.9233676791 | 1 | 0 |
| [9,] | 0.9983047247 0.0016953100 | 0 | 0 |
| [10,] | 0.1018455178 0.8981544375 | 1 | 0 |
| [11,] | 0.6893388033 0.3106612265 | 0 | 0 |
| [12,] | 0.0130585674 0.9869413972 | 1 | 0 |

**Figure 21: Training process with balanced images data**. The network was training for 80 epochs with batch size 32, dropout of 0.25 and loss of function binary cross entropy. The overall accuracy was 1 for training and 0.71 for test sets. It accurately predicts the class labels of all images as shown on confusion matrix table (B) and first five images (C) on training dataset but fail to predict all the validation sets (D and E).

# IV. CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Human embryo evaluation based on static morphological features such as TE, ICM and ZP is the standard IVF clinics around the world[4,39]. The subjective nature of this time-consuming process results in discrepant classifications among embryologist and clinics leading to fail in accurately predict embryo implantation and live birth potential[3,39,57,87]. Time lapse elimination of periodic transfer to microscopy assessment, stable embryo culture conditions for embryo development and morphokinetic properties, has alleviate the problem. Therefore, there are significant limitations even considering morphokinetics analysis. Likewise, many patients require multiple IVF to achieve pregnancy, making the selection of single embryo for transfer a critical challenge.

Few studies involving TLI of human embryos and deep learning have been done since the introduction of TLI in clinical election of single best embryo for intrauterine transfer. TLI give rise to images consistence in terms of light, size, quality and developmental timing records, which is particularly important when quantifying blastocyst expansion[147]. TensorFlow in tandem with the Keras library, become Deep learning more accessible for companies as well for individuals[148]. Deep learning allows the discover of structures in a large dataset using a back-propagation algorithm and to conduct small changes in its parameters in order to achieve the algorithm with the optimal representation of the dataset[100]. Henceforth, there is lack of automated methods to extract and quantify features associated to embryos outcomes in TLI data. To the best of our knowledge this is the first study in Portugal to develop an automated embryo classification system that should allow a more objective and accurate analyses and guide the clinicians in decision-making.

The current study presents a pipeline based on deep learning for TLI classification. The accuracy on the final validation dataset reached, in the best scenario, 86% (Table 6). Although the results presented here are preliminary, the framework allows clinicians and researchers, with no expertise in machine learning, to utilize deep learning to gather information about embryo quality. Furthermore, these results suggest that as much relevant clinical information as possible, including pregnancy rate, birth rate for viable or duration of pregnancy, should be stored for future use, levering the potential coupling with logistic regression. Only three known study to date reach such predictions capacity. Iwata and colleagues[148] using 118 images of human embryos obtained from high-resolution time-lapse cinematography could predictively determine good-quality embryos with 94% for training dataset and 70% for validation dataset. Khoshavi and colleagues[147], in a more robust study using 10,148 TLI of human embryos and

trained deep neural network Inception-V1 called STORK, found prediction accuracy of 98%, with 0,98 AUC. In a more recent study, Chen and co-workers[149], develop a CNN-based prediction model with three classification categories of blastocyst, ICM, and TE. They achieve an overall predictive accuracy of 75,36% using more than 170,000 embryo raw digital images from Asian populations. The CNN are based on ImageNet architecture. Our framework shows a promising improvement of accuracy performance, however it involves low number of embryos images from a single fertilization clinic and lack of independent studies. Inferior performance are achieved by Iwata and colleagues, which used same Keras libraries and image number approach. Compared to STORK and Chen´s framework, we used our proper designed CNN architecture, instead of pre-trained one, with consistent results. Also, we do not use common images segmentation techniques, instead we use a raw TLI, leading the CNN to access features and images patterns that embryologist and clinical practitioners are not able to access[147].

One of the greatest challenges associated with machine learning, including deep learning, is the prevention of overfitting - a condition in which the model cannot be applied to unknown data because it has been overly adjusted to the training data[88]. In the present study, the discrepancy between the training curve and validation curve suggests that overfitting occurred (Figure 16), most likely due to the small number of included images[110,141,148].

We explored methods to improve the generalization properties and to mitigate overfitting under the extremely limitation of an insufficient number of images, by augment artificially the image data via random transformations of existing images. The goal was exposing the model to more aspect of data and help it to generalize better, once it will not see the same image twice. The results do not show an improvement of CNN. Indeed, augmentation seem to worse the overfitting limitations. The results suggest this performance deterioration are due to heavily transformations that deprive useful information to the model and could act as distracting noise, increasing overfitting and generalization error[145]. This idea is observed when these arguments are turned false or eliminated (Figure 21 and 22) or when manually transformed images are added to training and test datasets (Figure 17). As suggested by Gardner and co-workers[39] in time lapse images embryos, features orientation, intensity and lateral/horizontal asymmetry sound to be important to the embryo morphology, so heavily augmentation technique might damage the image content and difficult features extraction by the network[145]. It appears that the augmentation techniques adopted have a large effect in class discrimination and network generalization performance, putting the transformations outside of the range and may not preserve the class specificity. In fact, with deformation on time lapse

images of blastocyst, in a heavily augmentation techniques on Keras, some important features directly related with strong clinical outcomes, such as full blastocoel cavity, ICM with tightly packed numerous cells and TE with many cells[39], should be degraded or even lost.

At some instances of results evaluation is seem that our model underestimates the labelled good quality class embryos (Figure 16, 22 and Table 6). Such class underestimation is due to class imbalance, since class 0 or poor-quality embryos, are twice of good quality class embryos. In binary classification problem with data samples from two groups, class imbalance occurs when one class contains significantly fewer samples than the other class. Should be the case, that model over-classify the poor quality class embryos, with more data, due its prior probability and underestimate minor good quality class group[146]. In fact, the cross-validation study denotes a clear over-estimation of poor quality embryos. Due to few numbers of good quality input images, the convolution operations do not extract all the hierarchical features associated with such class of embryos.

Here we propose and experimentally demonstrate the huge potential of deep learning combined with time-lapse imaging technology for embryo selection. Indeed, the presented framework represent a proof-of-concept that deep neural network discovers and exploits key biological features associated with day 5 time-lapsed embryos and should access time lapse images patterns that are not easily accessed by clinicians' practitioners. The approach presented in this thesis, suggest an immediate and wide applicability to improve time-lapsed embryos selection and over-pass the problem specific of embryo selection. The method is automatic, reproducible and objective in human blastocyst evaluation. Otherwise, the powerful learning capability of the deep learning approach to recognize and use biological features that are class-dependent from raw images, eliminate manual design and optimization of these features. Although not exhaustive, we present here a small contribution to the study of image augmentation processes in deep learning approaches in embryology images studies and medical imaging in general[147].

The pipeline and designed network were able to predict embryos quality with 86% of accuracy. Although the predictions accuracy and generalization properties must be improved, which can be achieve first by increase the number of image data, and second, with the data increment, update the deep learning properties. Also, coupled with new TLI data, there is a need to annotate relevant clinical information such pregnancy rate, duration of pregnancy, live

birth rate and born of health child rate. Therefore, this information should be protected by General Data Protection Regulation[c]. Otherwise, it will be interesting to investigate the black-box of features extraction of deep learning using LIME - Local Interpretable Model-agnostic Explanation, to explain the prediction of our binary classification.

[c]Article 9 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April.

# V. REFERENCES


1. Zegers-Hochschild, F. *et al.* International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) revised glossary of ART terminology, 2009. *Fertil. Steril.* 92 (5), 1520-1524 (2009).
2. Inhorn, M. C. & Patrizio, P. Infertility around the globe: New thinking on gender, reproductive technologies and global movements in the 21st century. *Human Reproduction Update* 21 (4), 411-426 (2014).
3. Pribenszky, C., Nilselid, A. M. & Montag, M. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. *Reproductive BioMedicine Online* 35 (5), 511-520 (2017).
4. Balaban, B. *et al.* Istanbul consensus workshop on embryo assessment: Proceedings of an expert meeting. *Reproductive BioMedicine Online* 26 (6), 1270-1283 (2011).
5. Payne, D., Flaherty, S. P., Barry, M. F. & Matthews, C. D. Preliminary observations on polar body extrusion and pronuclear formation in human oocytes using time-lapse video cinematography. *Hum. Reprod.* 12 (3), 532-541 (1997).
6. Storr, A., Venetis, C., Cooke, S., Kilani, S. & Ledger, W. Time-lapse algorithms and morphological selection of day-5 embryos for transfer: a preclinical validation study. *Fertil. Steril.* 109 (2), 276-283 (2018).
7. Armstrong, S., Bhide, P., Jordan, V., Pacey, A. & Farquhar, C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database of Systematic Reviews* 29 (5), CD011320 (2018).
8. Adamson, G. D. *et al.* International Committee for Monitoring Assisted Reproductive Technology: world report on assisted reproductive technology, 2011. *Fertil. Steril.* 110, 1067–1080 (2018).
9. Nelson, H. On the Reproduction of the Ascaris Mystax. 6. *Proc. R. Soc. London* (2006).
10. Bavister, B. D. Early history of *in vitro* fertilization. *Reproduction* 124(2), 181-196 (2002).
11. Pincus, G. & Enzmann, E. V. The Comparative Behavior of Mammalian Eggs *in vivo* and *in vitro*. *J. Exp. Med.* 62(5), 665-675 (1935).
12. Chang, M. C. Fertilization of rabbit ova *in vitro*. *Nature* 184 (Suppl 7), 466–467 (1959).
13. Yanagimachi, R. & Chang, M. C. Fertilization of hamster eggs *in vitro* [38]. *Nature* 200, 281-282 (1963).
14. Edwards, R. G., Bavister, B. D. & Steptoe, P. C. Early stages of fertilization *in vitro* of human oocytes matured *in vitro*. *Nature* 221, 632–635 (1969).
15. Edwards, R. G. Maturation *in vitro* of human ovarian oocytes. *Lancet.* 2(7419), 926-929 (1965).
16. Steptoe, P. C. & Edwards, R. G. Birth after the reimplantation of a human embryo. *Archives of Pathology and Laboratory Medicine* 2(8085), 366 (1992).
17. Edwards, R. G. Maturation *in vitro* of mouse, sheep, cow, pig, rhesus monkey and human ovarian oocytes. *Nature* 208 (5008), 349-351 (1965).
18. Steptoe, P. C. & Edwards, R. G. Laparoscopic recovery of preovulatory human oocytes after priming

of ovaries with gonatrophins. *Lancet.* 1 (7649), 683-689 (1970).

19. Steptoe, P. C. Lapraroscopy and ovulation. *Lancet.* 44 (1), 38-42. (1968).
20. Steptoe, P. C. & Edwards, R. G. Birth after the reimplantation of a human embryo. *Camden Fourth Ser.* 21, 17 (1978).
21. Filho, E. S., Noble, J. A. & Wells, D. Toward a method for automatic grading of microscope human embryo images. in *2010 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2010 - Proceedings* 1289–1292 (2010).
22. Middelburg, K. J., Heineman, M. J., Bos, A. F. & Hadders-Algra, M. Neuromotor, cognitive, language and behavioural outcome in children born following IVF or ICSI - A systematic review. *Human Reproduction Update*. (14(3), 219-231 2008).
23. Ludwig, A. K., Sutcliffe, A. G., Diedrich, K. & Ludwig, M. Post-neonatal health and development of children born after assisted reproduction: A systematic review of controlled studies. *European Journal of Obstetrics Gynecology and Reproductive Biology*. 127(1), 3-25 (2006).
24. Fauser, B. C. J. M. *et al.* Health outcomes of children born after IVF/ICSI: A review of current expert opinion and literature. *Reproductive BioMedicine Online*. 28(2), 162-182 (2014).
25. Bay, B. Fertility treatment: Long-term growth and mental development of the children. *Dan. Med. J.* 61(10), B4947 (2014).
26. Basatemur, E. & Sutcliffe, A. Follow-up of Children Born after ART. *Placenta.* 29 (Suppl B), 135-140 (2008).
27. Sadler, T. W. (Thomas W. & Langman, J. M. embryology. Langman's medical embryology. *J. Chem. Inf. Model.* (2006).
28. Niakan, K. K., Han, J., Pedersen, R. A., Simon, C. & Pera, R. A. R. Human pre-implantation embryo development. *Development*. 139(5),829-841 (2012).
29. Poli, M. *et al.* Characterization and quantification of proteins secreted by single human embryos prior to implantation. *EMBO Mol. Med.* 7(11),1465-1479 (2015).
30. T.W.Sadler. Langman's Medical Embryology 12th Edition. 1, 225–227 (2012).
31. Stillman, R. J., Richter, K. S. & Jones, H. W. Refuting a misguided campaign against the goal of single-embryo transfer and singleton birth in assisted reproduction. *Human Reproduction*. 28(10), 2599-2607 (2013).
32. Johnston, J., Gusmano, M. K. & Patrizio, P. Preterm births, multiples, and fertility treatment: Recommendations for changes to policy and clinical practices. *Fertil. Steril.* 102(1), 36-39 (2014).
33. De Neubourg, D. *et al.* Single top quality embryo transfer as a model for prediction of early pregnancy outcome. *Human Reproduction.* 19(6), 1476-1479 (2004).
34. McLernon, D. J. *et al.* Clinical effectiveness of elective single versus double embryo transfer: Meta-analysis of individual patient data from randomised trials. *BMJ*. 341, c6945 (2011).
35. Pribenszky, C. *et al.* Pregnancy achieved by transfer of a single blastocyst selected by time-lapse monitoring. *Reproductive BioMedicine Online.* 21(4), 533-536 (2010).
36. Wong, C. C. *et al.* Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat. Biotechnol.* 28, 1115–1121 (2010).

37. M.D., V. *et al.* Computer-automated time-lapse analysis test results correlate to clinical pregnancy and embryo implantation: A prospective, blinded, multi-center study. *Human Reproduction* 29, 62–63 (2014).
38. Filho, E. S. *et al.* A method for semi-automatic grading of human blastocyst microscope images. *Human Reproduction.* 27(9), 2641-2648 (2012).
39. Gardner, D. K. & Balaban, B. Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': Is looking good still important? *Molecular Human Reproduction* 22, 704–718 (2016).
40. Scott, L. A. & Smith, S. The successful use of pronuclear embryo transfers the day following oocyte retrieval. *Human Reproduction* 13(4), 1003-1013 (1998).
41. Tesarik, J. & Greco, E. The probability of abnormal preimplantation development can be predicted by a single static observation on pronuclear stage morphology. *Human Reproduction.* 14 (5), 1318–1323 (1999).
42. Scott, L. The morphology of human pronuclear embryos is positively related to blastocyst development and implantation. *Human Reproduction* 15(11), 2394-2403 (2000).
43. Scott, L. Pronuclear scoring as a predictor of embryo development. *Reproductive BioMedicine Online.* 6(2), 201-214 (2003).
44. Aguilar, J. *et al.* The human first cell cycle: Impact on implantation. *Reproductive BioMedicine Online.* 28(4), 475-484 (2014).
45. Giorgetti, C. *et al.* Implantation: Embryo score to predict implantation after in-vitro fertilization: Based on 957 single embryo transfers. *Human Reproduction.* 10(9), 2427-2431 (1995).
46. Fisch, J. D., Rodriguez, H., Ross, R., Overby, G. & Sher, G. The Graduated Embryo Score (GES) predicts blastocyst formation and pregnancy rate from cleavage-stage embryos. *Human Reproduction* . 16(9), 1970-1975 (2001).
47. Pirkevi, C. *et al.* Synchronicity of cleavage cycles predicts blastocyst formation and quality. *Human Reproduction.* 100 (Suppl 3), 313 (2013).
48. Rienzi, L. *et al.* Significance of morphological attributes of the early embryo. *Reproductive BioMedicine Online.* 10 (5), 669-681 (2005).
49. Holte, J. *et al.* Construction of an evidence-based integrated morphology cleavage embryo score for implantation potential of embryos scored and transferred on day 2 after oocyte retrieval. *Human Reproduction.* 22(2), 548-557 (2007).
50. Ziebe, S. *et al.* Embryo morphology or cleavage stage: How to select the best embryos for transfer after in-vitro fertilization. *Human Reproduction.* 12(7), 1545-1459 (1997).
51. Gardner, D. K., Lane, M. & Schoolcraft, W. B. Physiology and culture of the human blastocyst. in *Journal of Reproductive Immunology*. 55(1-2), 85-100 (2002).
52. Gardner, D. K. & Lane, M. Culture and selection of viable blastocysts: A feasible proposition for human IVF? *Human Reproduction Update*. 3(4), 367-382 (1997).
53. Gardner, D. K. *et al.* Single blastocyst transfer: A prospective randomized trial. *Fertil. Steril.* 81(3), 551-555 (2004).

54. Van Den Abbeel, E. *et al.* Association between blastocyst morphology and outcome of single-blastocyst transfer. *Reproductive BioMedicine Online.* 27(4),353-361 (2013).
55. Du, Q. Y. *et al.* Blastocoele expansion degree predicts live birth after single blastocyst transfer for fresh and vitrified/warmed single blastocyst transfer cycles. *Fertil. Steril.* 105(4), 910-919 (2016).
56. Zhao, J., Yan, Y., Huang, X., Sun, L. & Li, Y. Blastocoele expansion : an important parameter for predicting clinical success pregnancy after frozen-warmed blastocysts transfer. 2, 1–8 (2019).
57. Kovacs, P. Embryo selection: The role of time-lapse monitoring. *Reproductive Biology and Endocrinology* 12, 124 (2014).
58. Filho, E. S. *et al.* A method for semi-automatic grading of human blastocyst microscope images. *Human Reproduction.* 27, 2641–2648 (2012).
59. Singh, A., Au, J., Saeedi, P. & Havelock, J. Automatic segmentation of trophectoderm in microscopic images of human blastocysts. *IEEE Trans. Biomed. Eng.* 62, 382–393 (2015).
60. Blandau, R. J. & Rumery, R. E. The attachment cone of the guinea pig blastocyst as observed under time-lapse cinematography. *Fertil. Steril.* 8(6), 570-585 (1957).
61. Mio, Y. & Maeda, K. Time-lapse cinematography of dynamic changes occurring during in vitro development of human embryos. *Am. J. Obstet. Gynecol.* 199(6), 660.e1-5 (2008).
62. Wong, C. C. *et al.* Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat. Biotechnol.* 28, 1115–1121 (2010).
63. Verlinsky, Y. & Kuliev, A. Human preimplantation diagnosis: needs, efficiency and efficacy of genetic and chromosomal analysis. *Baillieres. Clin. Obstet. Gynaecol.* 8(1), 177-196 (1994).
64. Treff, N. R. & Scott, R. T. Methods for comprehensive chromosome screening of oocytes and embryos: Capabilities, limitations, and evidence of validity. *J. Assist. Reprod. Genet.* 29(5), 381–390 (2012).
65. Gardner, D. K., Lane, M., Stevens, J. & Schoolcraft, W. B. Noninvasive assessment of human embryo nutrient consumption as a measure of developmental potential. *Fertil. Steril.* 76(6),1175-1180 (2001).
66. Kovacs, P. *et al.* Non-invasive embryo evaluation and selection using time-lapse monitoring: results of a randomized controlled study. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 233, 58-63 (2018).
67. Montag, M. H. M., Pedersen, K. S. & Ramsing, N. B. Time-lapse imaging of embryo development: Using morphokinetic analysis to select viable embryos. *Cult. Media, Solut. Syst. Hum. ART* 211–234 (2014).
68. Unisense FertiliTech. EmbryoScope $^{TM}$ Time-lapse System: the vision to conceive. (2015).
69. Ecso Medical. MIRI$^{TM}$ Multiroom Incubator for IVF. 9010137_Art_Equipment_MIRI_Brochure_vH.
70. Faramarzi, A., Ali Khalili, M., Micara, G. & Agha-Rahimi, A. Revealing the secret life of pre-implantation embryos by time-lapse monitoring: A review. *Int J Reprod BioMed* 15, 257–264 (2017).
71. Cruz, M. *et al.* Timing of cell division in human cleavage-stage embryos is linked with blastocyst formation and quality. *Reproductive BioMedicine Online.* 25(4), 371-381 (2012).
72. Kirkegaard, K., Kesmodel, U. S., Hindkjær, J. J. & Ingerslev, H. J. Time-lapse parameters as predictors of blastocyst development and pregnancy outcome in embryos from good prognosis

patients: A prospective cohort study. *Human Reproduction.*. 28(10), 2643-51 (2013).

73. Chamayou, S. *et al.* The use of morphokinetic parameters to select all embryos with full capacity to implant. *J. Assist. Reprod. Genet.* 30(5), 703-710 (2013).
74. Meseguer, M. *et al.* The use of morphokinetics as a predictor of embryo implantation. *Human Reproduction.* 26(10), 2658-2671 (2011).
75. Chen, M., Wei, S., Hu, J., Yuan, J. & Liu, F. Does time-lapse imaging have favorable results for embryo incubation and selection compared with conventional methods in clinical *in vitro* fertilization? A meta-analysis and systematic review of randomized controlled trials. *PLoS One*. 12(6), e0178720 (2017).
76. Goodman, L. R., Goldberg, J., Falcone, T., Austin, C. & Desai, N. Does the addition of time-lapse morphokinetics in the selection of embryos for transfer improve pregnancy rates? A randomized controlled trial. *Fertil. Steril.* 105(2), 275-285 (2015).
77. Armstrong, S., Arroll, N., Cree, L. M., Jordan, V. & Farquhar, C. Time-lapse systems for embryo incubation and assessment in assisted reproduction ( Review ). *Cochrane Database Syst Rev* (2015).
78. 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(7), 1025-1030 (2015).
79. Motato, Y. *et al.* Morphokinetic analysis and embryonic prediction for blastocyst formation through an integrated time-lapse system. *Fertil. Steril.* 105(2), 376-834 (2016).
80. Ciray, H. N. *et al.* Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Human Reproduction*. 29(12), 2650-2660 (2014).
81. Meseguer, M. *et al.* The use of morphokinetics as a predictor of embryo implantation. *Hum. Reprod.* 26, 2658–2671 (2011).
82. Basile, N. *et al.* The use of morphokinetics as a predictor of implantation: A multicentric study to define and validate an algorithmfor embryo selection. *Human Reproduction* 30, 276–283 (2015).
83. Conaghan, J. *et al.* Improving embryo selection using a computer-automated time-lapse image analysis test plus day 3 morphology: Results from a prospective multicenter trial. *Fertil. Steril.* 100(2), 412-419 (2013).
84. VerMilyea, M. D. *et al.* Computer-automated time-lapse analysis results correlate with embryo implantation and clinical pregnancy: A blinded, multi-centre study. *Reproductive BioMedicine Online.* 29(6), 729-736 (2014).
85. Goodman, L. R., Goldberg, J., Falcone, T., Austin, C. & Desai, N. Does the addition of time-lapse morphokinetics in the selection of embryos for transfer improve pregnancy rates? A randomized controlled trial. *Fertil. Steril.* 105(2), 275-285 (2016).
86. Liu, Y., Chapple, V., Feenan, K., Roberts, P. & Matson, P. Time-lapse deselection model for human day 3 in vitro fertilization embryos: The combination of qualitative and quantitative measures of embryo growth. *Fertil. Steril.* 105(3), 656-662 (2016).
87. Petersen, B. M., Boel, M., Montag, M. & Gardner, D. K. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on

Day 3. *Human Reproduction*. 31(10), 2231-2244 (2016).

88. Chollet, F. & Allaire, J. J. Deep Learning with R. Manning Publications. MEAP Edition (2017).
89. Ghatak, A. Machine Learning with R. *Springer Nature Singapure* (2017).
90. Bischl, B. *et al.* Machine learning in R. *The Journal of Machine Learning Research* (2016).
91. Sugiyama, M. Introduction to Statistical Machine Learning. *Morgan Kaufman*. 1st Edition. (2015).
92. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science*. 349(6245), 255-260 (2015).
93. Kotsiantis, S. B. Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana)*. 31(3) (2007).
94. Smola, A. & Vishwanathan, S. V. N. Introduction to machine learning. *University Press, Cambrige* (2014).
95. Bengio, Y. Learning Deep Architectures for AI. *Found. Trends® Mach. Learn.* 2(1),1-127 (2009).
96. Zhu, X. & Goldberg, A. B. Semi-Supervised Learning Tutorial. *Synth. Lect. Artif. Intell. Mach. Learn.* (2009).
97. Olivier, C., Bernhard, S. & Alexander, Z. Introduction to Semi-Supervised Learning. In *Semi-Supervised Learning* (2013).
98. Gibney, E. DeepMind algorithm beats people at classic video games. *Nature.* 518(7540), 465-466 (2015).
99. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*. 34 (6), 26-38 (2017).
100. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature.* 521, 436–444 (2015).
101. Johnson, K. W. *et al.* Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*. 71 (23), 2668-2679 (2018).
102. Ciregan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 3642-3649 (2012).
103. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770-778 (2016).
104. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (6), 1137-1149 (2017).
105. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., … Kingsbury, B. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*. 29 (6), 82-97 (2012).
106. Graves, A., Mohamed, A. R. & Hinton, G. Speech recognition with deep recurrent neural networks. in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2013).
107. Liu, J., Chang, W.-C., Wu, Y. & Yang, Y. Deep Learning for Extreme Multi-label Text Classification. in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '1,* 115-124 (2017).

108. Collobert, R. & Weston, J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. in *Proc. of the 25th Int. Conf. on Machine Learning*. 160-167 (2008).
109. Jones, N. Computer science: The learning machines. *Nature.* 505, 146-148 (2014).
110. Chen, X. W. & Lin, X. Big data deep learning: Challenges and perspectives. *IEEE Access.* 2, 514-525 (2014).
111. Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. & Mougiakakou, S. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Trans. Med. Imaging*. 35 (5), 1207-1216 (2016).
112. Roth, H. R. *et al.* Anatomy-specific classification of medical images using deep convolutional nets. in *Proceedings - International Symposium on Biomedical Imaging*. 101-104 (2015).
113. Cao, Y. *et al.* Improving Tuberculosis Diagnostics Using Deep Learning and Mobile Health Technologies among Resource-Poor and Marginalized Communities. in *Proceedings - 2016 IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2016* (2016).
114. Van Grinsven, M. J. J. P., Van Ginneken, B., Hoyng, C. B., Theelen, T. & Sánchez, C. I. Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images. *IEEE Trans. Med. Imaging*. 35 (5), 1273-1284 (2016).
115. Choi, J. Y. *et al.* Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. *PLoS One.* 12(11), e0187336 (2017).
116. Liu, S. *et al.* Early diagnosis of Alzheimer's disease with deep learning. in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (2014).
117. Prasoon, A. *et al.* Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 16(Pt 2), 246-253 (2013).
118. Yoo, Y., Brosch, T., Traboulsee, A., Li, D. K. B. & Tam, R. Deep Learning of Image Features from Unlabeled Data for Multiple Sclerosis Lesion Segmentation. In *International Workshop on Machine Learning in Medical Imaging. MLMI 2014*. 8679, 117-124 (2014).
119. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*. 16(6), 321-332 (2015).
120. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838 (2015).
121. Paggi, J. M. & Bejerano, G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA.* 24(12), 1647-1658 (2018).
122. Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248 (2017).
123. Ravi, D. *et al.* Deep Learning for Health Informatics. *IEEE J. Biomed. Heal. Informatic*. 21(1), 4-21 (2017).
124. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document

recognition. *Proc. IEEE*. 86(11), 2278 - 2324 (1998).

125. Szegedy, C. *et al.* Going deeper with convolutions. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2015).
126. Glorot, X., Bordes, A. & Bengio, Y. Deep Sparse Rectifier Neural Networks. *JMLR W&C.* 15, 315-323 (2011).
127. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. abs/1502.03167 (2015).
128. Ian Goodfellow, Bengio, Y. & Courville, A. *Deep learning*. *Nature Methods* (2017).
129. Chollet, F. Keras: Deep Learning library for Theano and TensorFlow. *GitHub Repos.* (2015).
130. Chollet, F. Keras as a simplified interface to TensorFlow: Tutorial. *The Keras Blog* (2016).
131. Wongsuphasawat, K. *et al.* Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Trans. Vis. Comput. Graph.* 24 (1), 1-12 (2018).
132. Abadi, M. *et al.* TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning. *Proc 12th USENIX Conf. Oper. Syst. Des. Implement.* 265-283 (2016).
133. Goldsborough, P. A Tour of TensorFlow Proseminar Data Mining. *Arxiv* (2016).
134. Anastassiou, G. *et al.* TensorFlow Tutorial. *Neural Comput.* (2015).
135. Rampasek, L. & Goldenberg, A. TensorFlow: Biology's Gateway to Deep Learning? *Cell Syst.* 2 (1) 12-14 (2016).
136. Adolfsson, E., Porath, S. & Andershed, A. N. External validation of a time-lapse model; a retrospective study comparing embryo evaluation using a morphokinetic model to standard morphology with live birth as endpoint. *J. Bras. Reprod. Assist.* 22(3), 205-214 (2018).
137. Van Der Laak, J. A. W. M., Pahlplatz, M. M. M., Hanselaar, A. G. J. M. & De Wilde, P. C. M. Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy. *Cytometry.* 5(3), 2-8 (2000).
138. Vincent, L., Vincent, L. & Soille, P. Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 583–598 (1991).
139. Jo, Y. J. *et al.* Holographic deep learning for rapid optical screening of anthrax spores. *Sci. Adv.* 3(8), e1700606 (2017).
140. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic gradient descent. *ICLR Int. Conf. Learn. Represent.* (2015).
141. Eaton-rosen, Z. & Bragman, F. Improving Data Augmentation for Medical Image Segmentation. *Midl* (2018).
142. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30(7), 1145 -1159 (1997).
143. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 143(1), 29-36 (1982).
144. Sutskever, I., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. R. Dropout : A Simple Way to Prevent

Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1929−1958 (2014).

145. Hussain, Z., Gimenez, F., Yi, D. & Rubin, D. Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. *Annu. Symp. proceedings. AMIA Symp.* 979–984 (2017).
146. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data*. 6, 1-54 (2019).
147. Khosravi, P. *et al.* Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *npj Digit. Med.* 2 (21), 1-9 (2019).
148. Iwata, K. *et al.* Deep learning based on images of human embryos obtained from high-resolusion time-lapse cinematography for predicting good-quality embryos. *Fertil. Steril.* 110 (4), e213 (2018).
149. Chen, T.-J. *et al.* Using Deep Learning with Large Dataset of Microscope Images to Develop an Automated Embryo Grading System. *Fertil. Reprod.* 1(1), 51-56 (2019).