EXPLORING THE COMPLEXITIES OF AI-ASSISTED EMBRYO SELECTION: A COMPREHENSIVE REVIEW

LUCIA URCELAY GANZABAL

Thesis supervisor: DARIO GARCÍA GASULLA (Department of Computer Science)

Thesis co-supervisor: ÀTIA CORTÉS MARTINEZ

Degree: Master Degree in Artificial Intelligence

Thesis report

School of Engineering
Universitat Rovira i Virgili (URV)

Faculty of Mathematics
Universitat de Barcelona (UB)

Barcelona School of Informatics (FIB)
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

28/06/2023

Abstract

In-Vitro Fertilization is among the most widespread and successful treatments for infertility. One of its main challenges is the evaluation and selection of embryos for implantation, a process which suffers from large inter- and intra-observer variability. Due to the advancements in time-lapse imaging, Deep Learning (DL) methods are gaining attention to address this issue, raising both technical and ethical questions. The published works on the topic either fail to address the generality of the problem by focusing on a particular approach or compare different approaches in a misleading manner. In this master thesis, we present and compare different DL-based alternatives delving into technical characteristics, explainability aspects, ethical considerations and clinical applications. Moreover, we propose a set of guidelines for the development of an AI model for embryo selection based on the previous analysis of the literature. Our ultimate goal is to offer a better understanding of the complexities involved in this problem as a necessary first step while working in such a sensitive domain.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Infertility is a common reproductive health problem that affects millions of people worldwide, causing social, psychological, physical and economic distress to the ones seeking to conceive [1]. In the coming years infertility rates are projected to grow due to environmental and lifestyle factors [2, 3]. In vitro fertilization (IVF) technology is used to overcome infertility, it involves the fertilization of an oocyte with sperm in the laboratory, followed by the transfer of the resulting embryos into the patient's uterus. The main challenge of IVF is the selection of the embryos that will be either selected for implantation, cryopreserved (for later implantation) or discarded (if they exhibit undesirable features). This selection is to be performed during the early days after embryo insemination, typically between three and five days. During this time, embryos are monitored in time-lapse imaging incubators (TLI), facilitating uninterrupted embryo growth within stable culture conditions (Figure 1.1). This technology offers a dynamic perspective on in vitro embryonic development, augmenting the clinical effectiveness of IVF [4].

To assess embryo quality, embryologists evaluate different morphological characteristics depending on the embryo development phase. Early development (days one to three) focus on cell number, symmetry and fragmentation rate, while embryos reaching bastocyst stage (day five) are also assessed by their expansion grade and the appearance of the inner cell mass (ICM) and the Zona Pellucida (ZP), as well as to the trophectoderm cells (TE) [5]. Figure 1.2 illustrates a 3 day embryo and a 5 day embryo along their main morphological features. These features are currently the best available evidence regarding the quality of embryos, and represent the foundation of current development assessment guidelines such as Gardner's [6] or ASEBIR in Spain. However, these

Figure 1.1: Embryoscope. *Wings Embryoscope* by Dr. Jayesh Amin, CC BY-SA 3.0

approaches are limited by the subjective assessment of embryologists, which causes inter- and intra-observer variability, and restricts the success rates of IVF. According to [7], the national pregnancy rate in Spain via IVF in the year 2019 stood at 41.9%, while the birth rate stood at 30,7%.

Artificial Intelligence (AI), and specially Deep Learning (DL) due to its capacity for dealing with images, have recently been considered to assist in the embryo assessment and selection process. AI has the potential to facilitate and improve the process of embryo selection, increasing the implantation success rates, and reducing the chances of multiple pregnancies. AI can also mitigate inter- and intra-observer variability, making results more reproducible and consistent [9]. Finally, AI can help reduce the financial, physical and emotional burden on patients, by optimising the treatment plan and minimising the need for repeated cycles of IVF.

Nonetheless, several aspects of AI models for embryo selection remain unclear and challenging. One major problem is the lack of standardisation and openness of the procedures for training these kinds of systems, which hinders reproducibility and comparability of results across different models. A variety of performance metrics are used in the field, making it difficult to directly compare the effectiveness of different approaches. Furthermore, comparisons between studies are often made on different outcomes and data foundations, *e.g.*, using different patient demographics, unbalanced data, or sub-cohorts, further complicating the ability to draw meaningful conclusions from literature [10]. The absence of a framework makes difficult the assessment and comparison of differ-

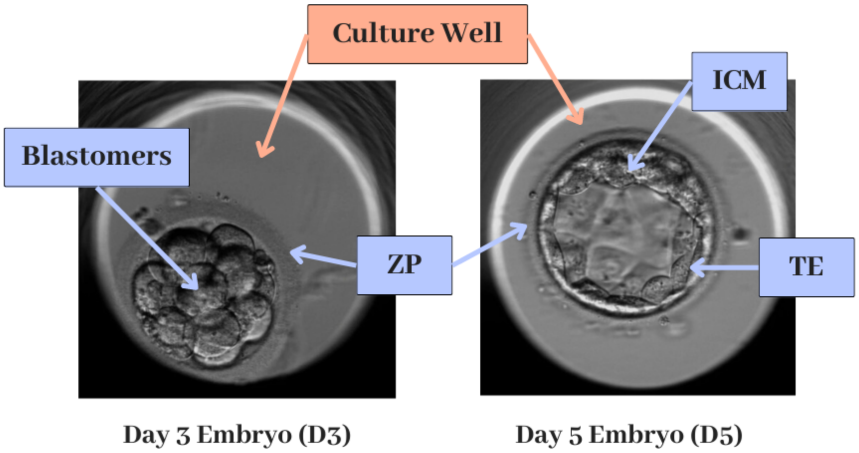

Figure 1.2: Morphological features of D3 and D5 embryos (blue) and culture well (orange). Edited images, originals from [8].

ent systems. Thus, establishing a common ground for reporting and comparing performance metrics is required.

Despite its importance and criticality, the potential bias of AI systems in embryo selection is rarely addressed in the literature. Biases in AI-assisted embryo selection can appear in many forms; while the model itself may present biases favoring or discriminating subpopulations (*e.g.*, patient age or fertilization method), biases may also be present in the use of these systems in a clinical context, over-relaying or subjectively interpreting recommendations. Moreover, the adoption of theses system is currently limited by their lack of explainability due to their opaque nature, which can have significant implications. Clinicians and patients need to understand the decision-making process behind the outputs of these models to establish trust in their recommendations. Assuring that the decisions of the model are explainable is a requisite for assuring trustworthiness and building acceptance in these systems.

The work done within this thesis is part of an interdisciplinary collaboration between the HPAI research group at the Barcelona Supercomputing Center (BSC) and expert embryologists from the Clinic Hospital in Barcelona. The preliminary goals of this collaboration encompass a comprehensive review and critique of all aspects related to AI-Assisted embryo selection as a necessary first step towards more ambitious goals. For that purpose, this thesis will produce a set of recommendations for the responsible and reliable development of such systems. These comprehend best practices for each phase of model development, including defining the intended use of the system, data pre-processing, and model design, as well as evaluation methodologies. The long-term goals of this collaboration involve the development of AI models for embryo selection based on the findings and recommendations produced in this thesis. The insights gained from the literature review and critical analysis presented in this work will heavily influence the development of the models, with the aim of building reliable and robust systems in which experts can find value.

Chapter 2

Background and Related Work

2.1 Embryo Selection in IVF

In vitro fertilization is an assisted reproductive technology that involves the manipulation of oocytes and sperm outside of the human body, in a controlled laboratory environment, to facilitate fertilization and the development of embryos. In this section, a review the IVF process and the data associated to it is carried out.

2.1.1 The IVF Process

The first step of the process is ovarian hyperstimulation, where the patient receives medication to stimulate the maturation of multiple eggs within the ovaries. This step aims to increase the number of eggs available for retrieval during the subsequent stages of the procedure. Once the eggs have sufficiently matured, they are retrieved from the ovaries in a procedure known as transvaginal oocyte retrieval. Simultaneously, the male partner or sperm donor provides a semen sample, which undergoes laboratory processing to isolate and prepare the sperm for fertilization.

After the eggs and sperm have been collected, fertilization can be achieved through two methods: conventional insemination, referred to as IVF in this project, or Intracytoplasmic Sperm Injection (ICSI). In conventional insemination, the eggs and sperm are combined in a culture dish, and the egg naturally selects a spermatozoon for fertilization. In ICSI, a single sperm is injected directly into the egg. It is important to note that the choice of insemination method can introduce visual differences in the resulting embryos, as the injection method may leave a visible tail on the embryo while conventional insemination not, but in turn can leave free spermatozoon in the culture dish. This variations important for the task at hand, as it can potentially introduce a visual bias when developing an AI model for embryo selection. In clinical practice, individual oocytes are typically inseminated in batches and assigned the same starting point timestamp for practical reasons. However, it is important to consider that there may be variations of up to 10 minutes in the recorded timestamps. This consideration is crucial since the time of development between different stages plays a significant role in the morphokinetic evaluation during clinical assessments.

Following fertilization, the resulting embryos are cultured in a TLI. During this culture period, cumulus cells, which are the tissue that surrounds and protects the egg, are manually removed. However, some remnants of cumulus cells may remain visible. Although they do not interfere with embryo development, it is important to consider this potential source of bias in the AI model. Yet other potential source of biases include the change or alternation of culture medium used during embryo development or the performance of genetic tests such as PGT-A.

After cultivation, embryos must be either selected for insemination, cryptopreserved for future cycles or discarded. The final step of the IVF process involves transferring the selected embryos into the patient's uterus. The number of embryos transferred depends on the patient's clinical condition an also in the legislation of the region on which the procedure is performed. For example, in Spain, no more than three embryos can be transferred to the patient in a single cycle [11]. Following embryo transfer, the patient undergoes monitoring to determine if a successful pregnancy has been achieved. In cases where pregnancy is not achieved, the couple may opt to repeat the IVF cycle.

This section provides an overview of the fundamental steps involved in the IVF process, highlighting key factors that may introduce potential biases. These factors encompass the choice of fertilization method and its timing, the selection of culture medium, the technique employed for embryo transfer, and the incorporation of genetic testing.

2.1.2 Embryo Development: Stages and Morphokinetic Features

Embryos undergo different stages during the incubation, as illustrated in Figure 2.1. During the initial stages of embryo culture, typically from day 1 until around day 3, embryos go through the cleavage stage, which corresponds to approximately 70 hours post-insemination (hpi). This stage is characterized by the division of cells, known as blastomeres, derived from the fertilized egg. Blastomeres form a compact mass called the morula. During days 2 and 3 of development, clinicians assess the quality of the embryos. This project focuses on ASEBIR grading methodology, which is the grading system employed by the Clinic Hospital in their clinical practice. Relevant features in this stage include:

- Number of cells and division rate
- Percentage and type of cellular fragmentation
- Blastomer size
- Visualization of nuclei and degree of multi-nucleation
- ZP appearance
- Degree of compactation
- Early adhesion

After the cleavage stage, the embryo progresses to the blastocyst stage, which typically lasts until 140 hpi, which corresponds to the fifth day. At this point, the embryo undergoes significant morphological changes, and the evaluation parameters shift accordingly. During the blastocyst stage, clinicians assess the following features:

- Expansion grade
- ICM appearance
- TE appearance

When to Transfer? Day 3 vs. Day 5

One of the most important factors to consider in IVF is the time of transfer. Most frequently, this happens at either day 3 (70 hpi, end of cleavage stage) or at day 5 (120 hpi, blastocyst stage). Embryo transfer to the uterus at Day 3 (D3) has its own set of advantages and disadvantages. One of the main advantages is that there is a larger number of embryos ready to be transferred at this stage. This increases the chances of finding viable embryos for transfer and potentially achieving a successful pregnancy. Additionally, a higher number of embryos can be frozen for future use, providing an advantage in subsequent cycles if needed.

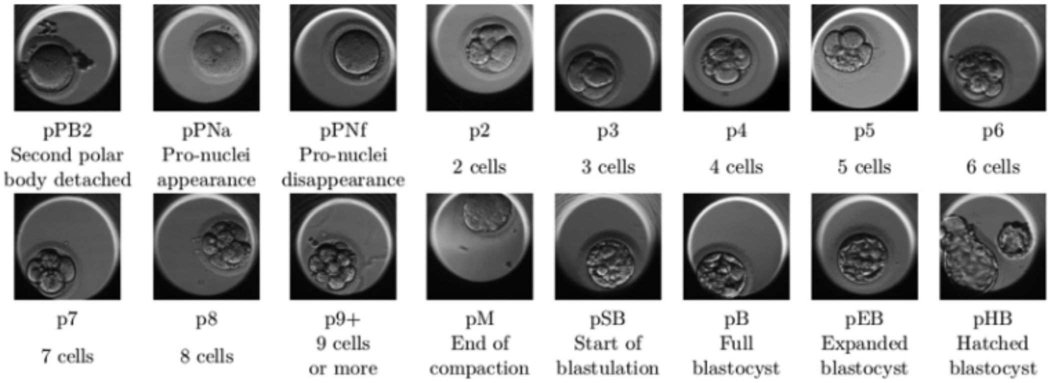

Figure 2.1: Development phases of the embryo during culture [8]. The images shows sequentially, from top left to bottom right, the main stages through which the embryo undergoes. Phases p8 and pB correspond to days 3 and 5 respectively.

However, there are some drawbacks to D3 transfer. The methods used for morphological scoring of pre-implantation embryos at this stage may not be sufficient to accurately select the embryo with the highest implantation potential. Even if the embryo with the best apparent quality is chosen, there is no definitive proof that it will continue to develop successfully after the transfer, particularly during the key stage of blastocyst formation. Furthermore, in order to increase the chances of success, multiple embryos are often transferred, which can lead to a higher likelihood of multiple pregnancies. This, in turn, carries potential risks for both the mother and the babies.

The alternative to D3 transfer, is to transfer once the blastocyst is successfully formed, that is on day 5 (D5), and offers its own set of advantages [6]. D5 transfer is associated with higher implantation rates, primarily because it allows for a better selection of embryos. At this point, the improved synchronization between the embryo and the endometrium closely resembles the natural conditions, which may increase the likelihood of successful implantation. With more information available, embryos can be selected more accurately based on their kinetics (*i.e.*, the evolution of visual features through time), and morphology, increasing the likelihood of selecting the embryo with the best implantation potential. Furthermore, transferring embryos at D5 increases pregnancy success rates since most embryos with abnormalities are unable to progress to the blastocyst stage. This selection process reduces the chances of transferring embryos with low implantation potential. As a result, and given the increased certainty on the embryo selection, success rates of single embryo transfer are increased, minimizing the risks associated with multiple pregnancies.

Let us now consider the limitations of with D5 transfer. It is possible for genetically normal embryos to fail to reach the blastocyst stage due to intrinsic embryo factors or clinical conditions. A high number of viable embryos may become nonviable by day 3 of development, and this proportion can be even more significant for older women [12]. This highlights the potential limitation of relying solely on blastocyst stage transfer, as it may result in the loss of embryos that could have been viable at an earlier stage. In conclusion, the choice between D3 and D5 transfer is not a trivial decision and solutions to make a more informed decision are needed.

2.1.3 Source and Nature of Embryo-related Data

Obtaining data and preparing datasets for the training of AI models in the field of embryo selection must be done with special care given its direct implications with human life. Embryo images used for training AI models are typically captured using microscopes contained in TLI devices. Due to the sensitivity of this data and the interest of scientists in having exclusive access to train their models, there is a limited availability of publicly accessible datasets in this domain.

The availability of public datasets is crucial, as it enables the development of new models and techniques. Secondly, public datasets serve as a benchmark for comparing the performance of different models. Without publicly available datasets, the ability to compare and evaluate the performance of AI models and to reproduce their results becomes limited.

As of today, there are a few public embryo datasets available, such as the dataset published by Gomez et al. [13], which provides data of 704 time-lapse (TL) embryo videos obtained by the EmbryoScope incubator at different time steps and focal planes, accounting for a total of 2,4M images. Early this year another dataset was been published by Kromp et al. [14]; it contains static blastocyst images from single focal planes as well as Gardner score annotations, making this dataset suitable for embryo quality grading. Although these datasets contribute to the open source community and can be highly useful, they lack essential metadata, necessary to design reliable experimentation. For example patient demographics and clinical settings related information. The absence of this information makes it difficult to assess and mitigate potential biases in the model evaluation process.

The issue of data availability is not the only problem associated to embryo related data. The question of whether data related to an embryo is considered personal data falls within the scope of the General Data Protection Regulation (GDPR) [15] in European Union (UE). According to the GDPR, "Personal data is any information that relates to an identified or identifiable living individual", where an identifiable person can be directly or indirectly identified using identifiers such as name, identification number, location data, or other specific factors related to their identity.

In the medical domain, compliance with the GDPR is of high importance as most patient related data is considered personal data and must be processed in accordance with the regulation. When it comes to embryology, different kinds of data are involved in the process, such as data related to the IVF patient (*e.g.*, age, location) and embryo images. However, there is some ambiguity regarding the classification of embryo images as personal data. Nonetheless, as a precautionary measure, this project has chosen to treat the images as personal data.

2.1.4 Embryo Images and Metadata

Embryo data typically consists of two main types: embryo images and associated metadata.

Images

In recent years, there has been a shift towards using TLIs for culturing embryos, which offer uninterrupted growth and stable conditions. This technology provides a dynamic perspective on in vitro embryonic development, allowing for the extraction of kinetic parameters in addition to standard morphological parameters.

TLIs capture snapshots of embryos at regular intervals, providing a temporal dimension to the data, *e.g.*, an image every 10 minutes during the 5-6 days of embryo development. Some time-lapse incubators even have the capability to capture images from different focal planes, *e.g.*, 7 different focal planes for each image. This feature is particularly useful for tasks such as cell counting, where overlapping cells may obscure others. However, analyzing large amounts of data corresponding to multiple focal planes and temporal frames becomes challenging.

In the context of AI models, different approaches can be taken with respect to how these embryo images are treated. Figure 2.2 illustrates different sample types for training the model. The simplest approach is to treat each image as

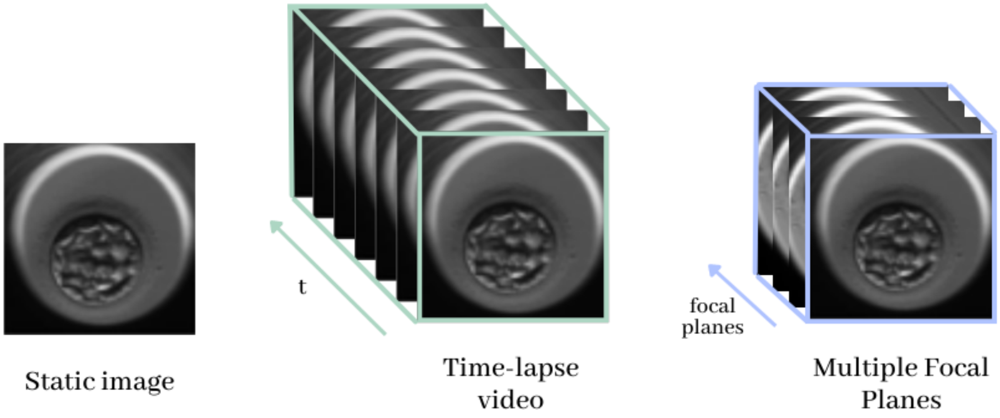

Figure 2.2: Sample types. From left to right: static image, time-lapse video of the same embryo at different time stamps, and collection of images of an embryo from different focal planes at the same time stamp. An additional and more complex sample type merges TL frames and focal planes in the same sample.

an individual sample, disregarding the temporal information and multiple focal planes. While this approach can increase data volume, it results in the loss of valuable kinetic information about the embryos as temporal evolution is lost. More ambitious solutions treat the entire TL video as a single sample, incorporating all the images captured over time and potentially combining them with data from different focal planes. This approach considers the complete developmental trajectory of the embryo. However, it also increases the complexity of the problem, requiring more advanced modeling techniques and computational resources. The definition of what constitutes a sample is crucial as it influences the intended use and performance of the AI model. Different approaches have their advantages and limitations, and the choice depends on the specific goals and requirements of the study or application.

Metadata

In addition to the embryo images, associated metadata can be collected, *e.g.*, clinical data, patient-related data and sample-related data. As mentioned earlier, this data serves not only as input for the AI model alongside the images but also as a means to account for potential biases that may exist within the dataset. It is known among clinicians that one of the most important factors to determine the success of an implantation is the age of the patient [12]. As a result, a model may demonstrate good performance for patients above the age of *e.g.*, 38 but not for younger patients. Having the relevant metadata helps identify and address such biases.

The metadata associated with embryo data includes information about the IVF clinical procedure itself, such as the insemination procedure, culture medium used, transfer protocol, the day of transfer, and the year of treatment. Patient-related data, such as maternal age, paternal age, and body mass index, also play a significant role in understanding the context of the data. Furthermore, sample-related data is collected, including the number of IVF cycles undergone, the number of embryos obtained per cycle, the number of images available per embryo, and the specific stage of development at which the embryos are imaged (e.g., Day 3 or Day 5).

2.2 Related work

Let us now summarize previous analyses of the AI-Assisted embryo selection field. These studies explore various aspects of this domain, and their limitations, as we will see by the end of the section, motivate our work and contributions.

Dimitriadis et al. [16] provide a broad overview of AI applications throughout the various stages of the IVF procedure, covering topics ranging from spermatozoa and oocyte analysis to models that target pronuclear, cleavage, and blastocyst stages. Conversely, there are other studies specifically focused on the stage of embryo selection in IVF. These can be categorized based on the specific task the studies perform.

For instance, Isa et al. [17] analyze papers that develop Machine Learning (ML) models for blastocyst grading, encompassing embryo classification models and embryo segmentation techniques. Others aim to encompass a wider array of tasks. Konstantinos et al. [18] discuss models for predicting clinical pregnancy, clinical pregnancy with fetal heartbeat, and ploidy status. The work of Louis et al. [19] explores approaches for embryo development phase annotation, including cell counting, detection and tracking, blastocyst formation and implantation potential prediction, and embryo grading and selection. Lundin et al. [20] focus on works related to TLI for embryo assessment and ploidy status determination. Zaninovic et al. [21] limit their study to ML models for automatic annotation, embryo grading and selection, and ploidy status prediction. Fernandez et al. [22] present a review that includes both traditional algorithms, such as Bayesian Networks and Support Vector Machines (SVM), as well as more modern approaches for embryo evaluation. Additionally, there are analysis, such as the one by Kim et al. [23], which provide a biological perspective, specifically discussing morphokinetic markers for predicting implantation potential and highlighting deep learning-based models developed for this purpose. Nonetheless, this studies are not without flaw, several issues can be identified within them.

Broad reviews and outdated literature

Firstly, some reviews aim to cover a broad range of tasks related to embryo selection without delving into the specific nuances of the problem. As a result, they may not provide comprehensive insights into the specific challenges and considerations that are relevant in the design of AI models. Furthermore, it is important to note that some of these reviews may be outdated, failing to capture recent advancements and breakthroughs in the field of DL methods specifically applied to embryo selection. Given the rapidly evolving nature of AI technologies, it is essential to rely on updated and relevant literature to gain accurate and reliable insights.

Lack of explainable AI

Importantly, while these reviews analyze the technical aspects of embryo selection, the topic of explainable AI in this context has not been adequately addressed. It is crucial to showcase the approaches that have been developed in explainability, highlighting their strengths and limitations, to guide future work. Explainable AI plays a critical role in ensuring that the decision-making process of the algorithm is transparent and understandable to clinicians and patients. Without explainability, there is a risk of lacking trust in the technology, which can ultimately hinder its widespread adoption.

Methodological shortcomings

Another significant drawback of these reviews is the methodology employed to compare the papers. Most of them separate studies by task, such as pregnancy prediction, and compare their performance using inconsistent metrics or experimental settings. This practice leads to inconsistent evaluation and can potentially mislead readers, as there is substantial variation in input data, embryo populations, and outcome measures across studies. For instance, Fernandez et al. [22] compare accuracy measures across different studies and datasets without considering the differences in embryo populations and label distributions within the test sets.

Lack of ethical considerations

Furthermore, an important concern regarding these reviews is the lack of attention given to the ethical aspects of AI-Assisted embryo selection, which are highly relevant in this application. The introduction of these models as Decision Support Systems (DSSs) in the clinical context remains largely unexplored. Consequently, crucial questions such as how to interpret the model's results, how to effectively communicate the decision to the patient or what impact can these systems have in the society are left unanswered. The implications of AI-assisted embryo selection in IVF must be thoroughly examined to ensure that the technology is developed and implemented in an ethical and responsible manner. While a few studies have studied these ethical considerations, such as the works [24, 25, 26], there is currently no comprehensive review that analyzes both the technical and ethical dimensions of this topic.

Given the drawbacks and limitations encountered in existing reviews, this project aims to provide a holistic overview of the topic, encompassing both technical and ethical aspects. Moreover, the objective is to compare the papers in a meaningful and non-misleading manner, addressing the gaps in previous comparative methodologies. By using this comprehensive approach, a more robust understanding of AI-assisted embryo selection can be achieved, serving as the foundation for the development of practical guidelines on designing reliable and trustworthy models in this project.

Chapter 3

Methodology

The preliminary goals of this project encompass a comprehensive review and critique of the aspects related to AI-Assisted embryo selection, with the aim of providing a set of recommendations for the development of robust and trustworthy AI-Assisted embryo selection models. This section outlines the methodology employed for the literature review process.

3.1 Search Strategy

To ensure comprehensive coverage of the literature, two search strategies were employed to identify and evaluate relevant studies. The first strategy focused on identifying papers that introduce novel DL models for embryo selection, while the second involved identifying papers that implement or assess explainability methods within their models. In terms of the timeframe, papers published from 2018 onwards were considered, as DL models in this field are relatively recent and rapidly evolving, with sigficand advances withing the last 5 years such as *EfficientNet*, *ConvNext*, *visual transformers* and *diffussion models*.

Three academic journal databases were used for all searches: PubMed, Scopus and IEEE. The common keywords used for the queries were: *embryo selection*, *artificial intelligence*, *deep learning* and *IVF*; for the search of explainability papers *explainability* and *XAI* were added. It should be noted that, although papers on explainability can be considered as a subset of papers on DL models, different results were obtained by using different key words in the queries. After searching in all the databases, 171 papers were identified in total for screening.

3.2 Study Selection

For papers proposing a DL model, this review has focused on three specific tasks: *Blastocyst Formation Prediction*, *Pregnancy Prediction*, and *Live Birth Prediction*. As a result, the in-depth analysis of other tasks related to AI-assisted embryo selection (*e.g.*, embryo segmentation, development stage identification) has been deferred to future work. In the case of explainability related papers, the inclusion was limited to works that applied any form of XAI method to their model, introduced a novel XAI method, or conducted comparisons between different XAI methods.

It is important to note that the literature review conducted for this study encompasses articles published up until June 2023. Finally, out of the 171 papers identified in the search, after duplicate removal, abstract screening and review of additional papers obtained from the references of selected studies, a total of 28 were fully reviewed and included in the study.

3.3 Information Extraction

As seen in the previous chapters, there are many variables involved in the design of AI models of embryo selection. One of the objectives of the review is to analyze and compare studies based on the consideration and selection of these variables. To provide methodological consistency, a series of parameters were defined to be annotated during the review of each paper.

1. Data related information: This included number of embryos considered, developmental stages of the embryos, fertilization method employed, use of fresh or frozen embryos, ploidy status of the embryos, hours post implantation of embryos in the images, frequency of image acquisition, image resolution, utilization of discarded embryos, labels and their distribution, number of images or videos utilized, number of patients involved, number of IVF cycles included, number of participating clinics or hospitals, study design (retrospective or prospective), and the imaging device employed TLI, microscope). Additionally, information regarding the number and types of focal planes used was also extracted.

2. Model related information: This included the specific task for which the model was designed (blastocyst formation prediction, pregnancy prediction or live birth prediction), the objective of the model (whether it focused on embryo-level prediction or patient-level embryo ranking), whether the

model incorporated explainability techniques, the type of explainable XAI method used if applicable, and the architecture of the model employed.

3. Evaluation related information: This involved the metrics used to evaluate the model's performance, the achieved performance results, whether subpopulation analysis was conducted to assess potential bias, the type of validation employed, evaluation of predictive performance over time, clinical assessment of the model, and whether an ethical evaluation was conducted.

The information extraction process consisted of two distinct phases. A first screening was centered in studies containing explainability methods, and was done in collaboration with other members of the HPAI research group, who reviewed XAI aspects from up to 25% of all found papers. Subsequently, a comprehensive and detailed review of all 28 papers was conducted solely by the author of this thesis. By extracting and analyzing these characteristics, along with the inclusion of papers focusing on ethical considerations, a comprehensive and holistic analysis of the current state of AI-assisted embryo selection in IVF was conducted.

To facilitate a comprehensive understanding and comparison of AI models, we adopt and extend the framework proposed by Kragh et al. [10], which utilizes a population-outcome scheme to characterize AI models in embryo evaluation based on their data foundation. In our extension, we include additional fields such hours post implantation, time intervals between image acquisition, use of discarded embryos, and labels and their distribution. This comprehensive framework allows for a better comparison between papers. Tables 3.1, 3.2, 3.3 and 3.4 show the data foundation of the different studies grouped by task. Symbol "-" was used to express that the information was not provided in that study, which in turn indicates the lack of transparency and reproducibility of these studies.

Ref.	N. of Embryos	Fert. Method	Fresh/ Frozen	Hours post Implantation	Image Frequency	Used Discarded Embryos	Labels & Distribution
[27]	3,300	-	-	-	5 min[1]	-	Blastocyst, non-blastocyst
[28]	2,898	-	-	0-168 h	6 h	-	Blastocyst (56.38%), non-blastocyst (43.62%)
[29]	12,912[2]	IVF, ICSI	-	-	5-10 min[3]	No	Blastocyst (n=5061), non-blastocyst (n=3285)
[30]	>6,200	ICSI	-	-	Unknown[4]	Unknown[5]	Blastocyst (50%), non-blastocyst (50%)[6]

Table 3.1: Data Foundation Blastocyst Formation Prediction Task

[1] After, key-frame selection model reduces frame number.
[2] Initially 26113, after D3 transferred/cryopreserved/discarded embryos elimination 12912 remained, not all were used for training.
[3] Frequency of the imaging device, unknown if it is the same frequency of input frames.
[4] Frequency seems to be different in different modules of the network.
[5] Embryos that underwent genetic testing were discarded.
[6] Balanced dataset is reported so 50% is inferred.

Ref.	N. of Embryos	Fert. Method	Fresh/ Frozen	Hours post Implantation	Image Frequency	Used Discarded Embryos	Labels & Distribution
[31]	181,428	ICSI, IVF	Fresh, Frozen	20-148 hpi, 20-84 hpi[2]	1 h	Yes[1]	KID+ (50%), KID- (25%), discarded (25%)[3]
[32]	9,359[4]	IVF	Fresh	-	Static images	Unknown[5]	KID+ (50%), KID- (50%),
[33]	115,832	ICSI, IVF	Fresh, Frozen	12-140 hpi	1 h	Yes	KID+ (50%), KID- (10%), discarded (40%)
[34]	17,984	ICSI, IVF	-	-	Static images	No	KID+, KID-
[29]	12,912	IVF, ICSI	-	-	5-10 min	No	Usable blastocysts (n=2,922), unusable blastocysts (n=1,356)[6]
[35]	946	ICSI	Fresh	-	Static images	Yes	Good prognosis, bad prognosis[7]

Table 3.2: Data Foundation Pregnancy Prediction Task (Part 1)

[1] Genetic testing was performed only to selected embryos.

[2] First time interval corresponds to Day 5+ model and second to Day 2/3 model. Embryos cultivated for less than 36 hpi and between 84-108 hpi were excluded.

[3] This data corresponds to Day 5+ model. Day 2/3 employs a more complex data split mechanism.

[4] Number of images, the number of embryos is not disclosed.

[5] Images were excluded if taken after biopsy for PGT-A or cryopreservation.

[6] A usable blastocyst means to have been chosen for transfer or vitrication. Reported data account for training set.

[7] A good prognosis is defined as either having a report of euploidy after PGT-A or a positive beta-hCG result.

Ref.	N. of Embryos	Fert. Method	Fresh/ Frozen	Hours post Implantation	Image Frequency	Used Discarded Embryos	Labels & Distribution
[36]	310	-	Fresh	113 hpi	Static images	-	KID+, KID-
[30]	>5,500	ICSI	-	-	Unknown	No	KID+ (50%), KID (50%)[3]
[37]	272	-	-	-	-	-	KID+ (n=216), KID- (n=56)
[38]	8,886	ICSI	Fresh[4]	-	Static images	Yes[4]	KID+ (50%), KID- (50%)[5]
[39]	8,836	-	Fresh, Frozen	-	-	No[6]	KID+ (n=694), KID- (8,142)[7]
[40]	344	-	-	-	Static images	-	KID+ (n=258), KID- (n=86)[8]

Table 3.3: Data Foundation Pregnancy Prediction Task (Part 2)

[1] A usable blastocyst means to have been chosen for transfer or vitrication. Reported data account for training set.
[2] A good prognosis is defined as either having a report of euploidy after PGT-A or a positive beta-hCG result.
[3] A balanced dataset is reported so 50% is inferred.
[4] Images were only accepted if they were taken prior to PGS biopsy or freezing.
[5] A balanced dataset is reported so 50% is inferred.
[6] Embryos that underwent embryo biopsy for pre-implantation genetic testing were included.
[7] Highly unbalanced dataset. A large proportion of predicted non-viable embryos were never actually transferred.
[8] Data augmentation was performed in the negative class to balance the dataset.

Ref.	N. of Embryos	Fert. Method	Fresh/ Frozen	Hours post Implantation	Image Frequency	Used Discarded Embryos	Labels & Distribution
[41]	15,434	-	Fresh, frozen	105-125 hpi	-	Yes	Live birth (50%), No live birth (50%)[1]
[34]	1,358	ICSI, IVF	-	-	Static images	No	Live birth, no live birth
[42]	470	ICSI, IVF	Fresh, frozen	-	Static images	Unknown[2]	Live birth, no live birth[3]
[43]	4,104	IVF	Fresh, frozen	115, 139 hpi	Static images	-	Live birth (38.7%), no live birth (61.3%)
[44]	263	IVF	-	-	-	No	Live birth, no live birth
[45]	5,691	IVF	Fresh, frozen	115, 139 hpi	Static images	-	Live birth (27.9%), no live birth (72.1%)

Table 3.4: Data Foundation Live Birth Prediction Task

[1] The dataset was initially unbalanced, it was balanced oversampling positive class.
[2] No PGT-A testing was done. Poor quality embryos were included for training.
[3] Unbalanced dataset. 5-fold cross-validation was performed with 18 positive embryos and 76 negative ones in each fold.

Chapter 4

Results

This section encompasses the presentation and discussion of the selected papers. Initially, the papers are organized and discussed based on the three distinct tasks identified in this study: *Blastocyst Formation Prediction*, *Pregnancy Prediction* and *Live Birth Prediction* (Figure 4.1). Subsequently, the focus shifts to the papers that incorporate an XAI component, where they are presented and analyzed in detail.

4.1 Tasks in AI-Assisted Embryo Selection

AI models of embryo selection can be categorized based on the specific task they aim to optimize. This task, in turn, depends on the specific time or stage of embryo development at which the selection is performed.

One common task is *Embryo Quality Grading*, which assess the quality of embryos based on their observable morphokinetic features. This assessment can be conducted at any stage during the culture period of the embryo. The commonly used ground truth for this task involves quality annotations made by experts during the embryo's development, following morphological annotation guidelines such as Gardner's or ASEBIR. However, it is important to note that the grading process is subjective and can vary among specialists, leading to potential inter- and intra-observer variability in the ground truth annotations. This approach is commonly adopted when Known Implantation Data (KID) is unavailable.

Another task related to embryos in cleave stage, at D3, is *Blastocyst Formation Prediction*. The objective of this task is to identify which embryos, at D3 of development, have the potential to progress and reach the blastocyst stage by

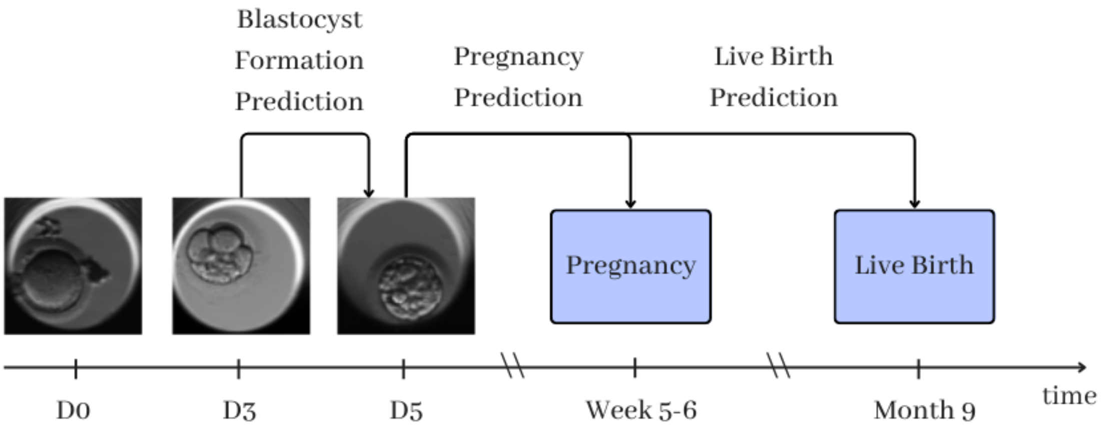

Figure 4.1: Evolution of an embryo and associated tasks.

D5, as elaborated upon in Section 2.1.2. This task will be explored further in the subsequent subsection.

When KID data is available, more ambitious tasks can be undertaken in the selection of embryos, which are based on their potential for successful implantation in the uterus and, ultimately, live birth. These are other of the two central tasks that occupy this study, *Pregnancy Prediction* and *Live Birth Prediction*, and will be discussed in depth in the following subsections.

Embryo selection can extend beyond considering only morphokinetic parameters related to their appearance. An optional practice is to perform genetic tests such as PGT-A on embryos to analyze their ploidy status, , which refers to having the correct number of chromosomes. Embryos that are aneuploid, meaning they have abnormal chromosome numbers, are less likely to develop into a healthy fetus and may result in miscarriage or birth defects. However, it is worth noting that methods like PGT-A are invasive and carry a risk of embryo damage or loss. As a result, researchers have developed DL models capable of performing *Ploidy Prediction*. These models aim to non-invasively predict the ploidy status of embryos, offering a potential alternative to invasive procedures like PGT-A.

In conjunction with the previous tasks, there are auxiliary tasks that complement the previous models. The *Development Stage Identification* of the embryo is necessary in automated labelling in TLIs. Automated tools are required for accurately measuring the timing parameters of multiple embryos in order to track the duration of their development stages. Such tools need to provide precise outcomes and be able to handle various challenges, including deforming cell shapes, poor visual features, and similarities between embryos at different stages. Finally, *Embryo Segmentation* can also be beneficial to the above tasks. Specifically, models have been developed to segment embryos by identifying their main biological structures. This segmentation allows for the extraction of parameters related to specific parts of the embryo or enables the training of models that specifically focus on processing these parts.

4.1.1 Blastocyst Formation Prediction Task

As explained in section 2.1.2, the decision of transferring the embryo on D3 or on D5 is not trivial, as both options present advantages and drawbacks. Developing models which are able to predict which embryos at D3 will develop into blastocysts by D5 is valuable for selecting the embryos with the greatest potential, specially in the case of older patients where development until D5 can result in arrested development. By accurately identifying these embryos, clinicians could potentially increase success rates while minimising risks associated with multiple pregnancies.

Table 4.1 contains the identified papers in the literature that perform *Blastocyst Formation Prediction* task. The four studies found are relatively recent, comprising years 2020-2022. All of them use TL videos for the development of the models, which allows them to capture the kinetic information from day 1 to day 3. Chen et al. [27] propose a framework that adaptively selects informative frames to predict blastocyst formation using TL videos at the cleavage stage on day 3. To complement it, another network generates predictions using the morphokinetics features of the selected frames. Xie et al. [28] develop a so-called multi-focus selection network (AMSNet) which includes an attention mechanism to exploit the features of TL images captured at multiple focal planes, as well as a temporal feature channel shift operation which enables it memory capability over TL videos.

Liao et al. [29] build a multi-module network. It is composed by an LSTM-based temporal stream model which, in turn, is combined with the output of a cell-counting module, and a spatial stream module which captures morphological features. Kan-Tor et al. [30] propose a two objective model: blastulation prediction and implantation potential prediction. In first place, TL images from each embryo are divided in 5-frame long "packets" and a network is trained to

Ref.	Input	Embryo Population	Clinician vs. Model	Metrics
[27]	TL video, frame number	D1-D3	-	Accuracy, sensitivity, specificity, PPV, NPV, F1, AUC
[28]	TL video, focal planes	D1-D7	-	Accuracy, AUC, ROC
[29]	TL video, frame number, cell count	D1-D3	-	Accuracy, sensitivity, specificity, PPV, NPV, AUC, ROC
[30]	TL video	D3	-	Sensitivity, PPV, AUC

Table 4.1: Studies on *Blastocyst Formation Prediction* Task. *Input* refers to the type of data used to train the model, *Embryo Population* to the kind of embryos that compose the dataset, *Clinician vs. Model* reports whether the performance the model has been validated by experts or compared to their performance, and *Metrics*, to the evaluation metrics that have been employed.

identify the time window to which the frames belong. Then a random forest and a logistic regression are trained with the previous output to predict blastulation and implantation potential.

4.1.2 Pregnancy Prediction Task

Pregnancy Prediction is a task that has gained significant attention in the literature, which is in turn more ambitious than the previous one. The development of AI models for pregnancy prediction from embryo images relies on the availability of known implantation data, in other words, if the embryo resulted or not in a successful pregnancy upon transfer. However, obtaining KID may be challenging as it requires continuous monitoring and data collection of the embryo after it has been implanted in the patient's uterus. Consequently, KID may be scarcer compared to annotations by experts on embryo quality.

Pregnancy can be measured using different endpoints, with fetal heart pregnancy being a common one. Fetal heart pregnancy refers to the detection of fetal heartbeats in the uterus through ultrasound, and it can be detected from week 5 onward. Other pregnancy indicators include the measurement of Human chorionic gonadotropin (hCG), a hormone produced by the placenta during pregnancy. HCG levels typically rise after conception and continue to increase until about 10 weeks into pregnancy. Additionally, ultrasound visualization of the gestational sacs can serve as an indicator of pregnancy.

These pregnancy indicators provide the basis for developing AI models to predict the likelihood of pregnancy. Among them, fetal heart pregnancy is the most commonly utilized. However, it is important to note that the successful implantation of a good-quality embryo does not guarantee a successful pregnancy. Various factors, such as the patient's age, pregnancy history, endometrial thickness, or progesterone levels, can influence the outcome and act as confounders in the model's predictions. In fact, according to [35], over 200 confounders exist affecting outcomes in assisted reproduction; as a result, it is unrealistic to expect that any embryo assessment can currently guarantee 100% prediction of a successful outcome. Therefore, the predictions made by the AI model should be understood as a likelihood of implantation rather than an assurance of pregnancy success. Table 4.2 contains the identified papers in the literature that perform *Pregnancy Prediction* task.

Lassen et al. [31] have published the latest work in implantation potential prediction measured by fetal heartbeat. It is worth noting that the authors are affiliated with Vitrolife, a prominent manufacturer of TLIs. They have developed the model with the largest and most diverse dataset until the moment, with data from 181,428 embryos from 22 different IVF clinics. Discarded embryos are included on the dataset, meaning that no previous evaluation by experts is required for the use of the models. Two different models have been developed, one for prediction of implantation at D2/3 and another for prediction at D5+. While the architecture of each of the models is slightly different, both are based on a combination of 3D-CNNs and logistic regression operations.

One year before, same authors from [31] (Lassen 2023) published the work [33] (Berntsen 2022), a model developed in a similar style which accomplishes the same task. The main difference lays in the design of the architecture of the DL model which, in addition to using a 3D-CNN to process the sequences of TL videos, they train a bidirectional LSTM which is able to capture the temporal information within the frames.

Ref.	Outcome	Input	Embryo Population	Clinician vs. Model	Metrics
[46]	Fetal heartbeat	TL videos	D1-D2; D1-D3; D1-D5+	-	AUC
[32]	Fetal heatbeat	Static Image	D5	✓	Accuracy, TPR, TNR, AUC, TTP
[33]	Fetal heartbeat	TL video	D0-D5	-	AUC
[34]	Fetal heartbeat, hCG	Static image, clinical data	D5	-	Accuracy, sensitivity, specificity, PPV, F1, AUC
[29]	Gestational sacs	TL video, frame number, cell count	D1-D3	-	Accuracy, sensitivity, specificity, PPV, NPV, AUC, ROC
[35]	Ploidy, hCG	Static image, patient age, blastocyst age, lab settings	D5,D6	✓	Accuracy, sensitivity, specificity, PPV, AUC, NDGG
[36]	"Pregnancy"	Static image	D5	✓	Accuracy, AUC
[30]	Fetal heartbeat, gestational sacs	TL video	D3, D5	-	Sensitivity, PPV, AUC
[37]	Fetal heartbeat, gestational sacs	TL video	-	✓	PPV, NPV, AUC
[38]	Fetal heartbeat	Static image	D5	✓	Accuracy, sensitivity, specificity
[39]	Fetal heartbeat	TL video	D5	-	AUC
[40]	"Pregnancy"	Static Image	-	-	Accuracy, sensitivity, specificity

Table 4.2: Studies on *Pregnancy Prediction* Task. *Outcome* refers to the variable measured to report pregnancy, *Clinician vs. Model* reports whether the performance the model has been validated by experts or compared to their performance.

Diakiw et al. [32] follow a different approach, where the aim is to develop models which perform embryo ranking within simulated cohorts instead of providing individual predictions of the embryos. To provide further insights, they compare the ranking provided by their model from the ranking resulting from traditional Gardner scoring system. Enatsu et al. [34] develop two kinds of models, one is a ResNet18-based model trained solely on static images and another is an ensemble model, also based in ResNet18 and a Random Forest Classifier, trained on both static images and clinical data. Results show that higher AUC is reached by the ensemble model, even trained with less data points than the former one.

The work of Liao et al. [29] is explained in *Blastocyst Formation Prediction* task, as it is its main goal. Nevertheless, as they also have implantation data for some of the embryos, they extend their model to predict implantation potential. It should be noted that due to the scarcity of KID, only the validation test set is composed by KID embryos. Chavez-Badiola et al. [35] propose a model for ploidy an implantation prediction composed by two modules. The first is designed to extract texture patterns from the images by applying a series of convolutions and then, segmenting the images into ROIs and extracting predictor features of embryo viability from each region. The second module is designed to rank embryos based on the scoring obtained form a DNN trained with previously obtained image-based features and metadata of each embryo.

While the main task of the work by Bormann et al. [36] is to discriminate blastocysts vs non-blastocysts in D5 in order to rank them within patient cohorts, the secondary objective is to predict implantation potential. To accomplish the latter task, they train a Xception-based Deep Neural Network (DNN) using KID data from 310 static images. Silver et al. [37] train a CNN autoencoder on individual frames of TL videos of unknown stage embryos in order to extract features; next, these features are used as input of an LSTM trained with 10 cross-validation which predicts implantation potential. Ver-Mileya et al. [38] present Life Whisperer model, an ensemble of eight different CNN-based models with a voting strategy on top. The models, which are based on different architectures such as ResNet, DenseNet and others, are trained on D5 static images of embryos. Results show an improved performance on accuracy of 30.8% when compared to the performance of embryologists in the same test set.

Tran et al. [39] published in 2019 *IVY* model, which predict fetal heartbeat via TL videos. Nevertheless, the performance of this model has been criticised due to the highly unbalanced dataset that they used for training. Cao 2018 [40] published in 2018 the first known approach for pregnancy prediction using a small dataset of static images; the model is based on a custom 10 layer DNN.

4.1.3 Live Birth Prediction Task

The next task to be addressed is *Live Birth Prediction*. This task can be seen as an extension of *Pregnancy Prediction*. While both tasks make use of known implantation data, live birth prediction is even more ambitious due to the multitude of factors that influence the successful birth of a child. Furthermore, failed implantation, miscarriage or embryo development failure can cause cost and time loss, and bring negative psychological outcome to the patient [43]. Table 4.3 shows the identified papers in the literature that perform *Live Birth Prediction* task.

Huang et al. [41] propose a custom CNN with residual connections to predict the live birth from TL videos of blastocyst stage embryos, although it remains unclear whether they process each frame individually or consider the TL video as a whole input. In a similar manner to Lassen et al. [31], they use discarded embryos and pseudo-label them as the negative class of "non live birth", this induces a source of bias since the label is inferred. Furthermore, they add that results of predicting live birth from cleavage stage embryos (D3) were not satisfactory. Sawada et al. [42] develop *Attention Branch Network* model, a 2D-CNN composed of two modules: a classifier and an attention module which outputs attention maps later used for visualization purposes. Although they have TL videos of variable length of 470 embryos, they process each frame individually and compute a weighted average for the final prediction at embryo level.

Miyagi et al. (2019) [45] develop a model for predicting live birth from static embryo images of blastocyst stage. On year later (2020), the same authors developed an enhanced version of the model [43] which, in addition to being trained with images, is trained with conventional evaluation parameters comprising patient data, clinical data and morphological embryo related data. Both types of data are combined by concatenating the feature maps of the images resulting from the CNN with the output of univariate regression functions for each variable of the metadata. When comparing the latest model's performance to the only-image model, they report that when the patient's age is less than 39 years old, the combined model outperforms the only-image one.

Silva-Rodriguez et al. [44] propose a model that, instead of predicting the probability of implantation from images, morphokinetic features are extracted in order to use them to train a Random Forest Classifier. The feature extraction module is composed of two methods: first, a CNN is trained to predict the num-

Ref.	Outcome	Input	Embryo Population	Clinician vs. Model	Metrics
[41]	Live birth	TL video	D5	-	AUC
[34]	Pregnancy, Live birth	Static image	D5	-	Accuracy, sensitivity, specificity, PPV, F1, AUC
[42]	Live birth	Static image	D0-D3, D0-D5	✓	Sensitivity, specificity, PPV, NPV, AUC
[47]	Live birth	-	D5	-	Accuracy, PPV, NPV, AUC
[43]	Live birth	Static image, annotations, patient data (age, BMI...)	D5/D6	-	Accuracy, sensitivity, specificity, informedness, PPV, NPV, AUC
[44]	Live birth	TL video	D0-D4	-	Accuracy
[45]	Live birth	Static image	D5-D6	-	Accuracy, sensitivity, specificity, AUC
[48]	Live birth	Static image, morphokinetic parameters	D5	-	Accuracy, AUC

Table 4.3: Studies on *Live Birth Prediction* Task. *Clinician vs. Model* reports whether the performance the model has been validated by experts or compared to their performance.

ber of cells in each frame of the TL video, then, temporal information is obtained by calculating intensity changes from on frame to another, which, according to them, is related with cell division.

4.2 Explainable Artificial Intelligence in Embryo Selection

The use of XAI on embryo images is reviewed in this section. It is worth noting that XAI is critical to ensure that the decision-making process of an algorithm is as transparent and understandable to clinicians and patients as possible. The lack of explainability can lead to a lack of trust in the technology, which ultimately hinders its adoption. Within the domain of images, explainable AI methods are often feature scorers, representing their output as saliency maps on the input image. Figure 4.2 shows saliency maps of two different methods applied in the same embryo image. Moreover, XAI techniques can also be used to investigate feature relevance of metadata associeted to images. By nature, XAI methods for DL are approximations to the model's real behavior. Given the limited availability of papers that implement XAI in the aforementioned tasks, the scope of the review was expanded to include other tasks within the field of embryo selection. These additional tasks include *Embryo Quality Grading*, *Embryo Development Stage Identification*, *Ploidy Prediction* and *Embryo Segmentation*, which have been introduced in section .

Table 4.4 shows studies found in the literature which integrate a XAI component. We can differentiate between those applied after the model is trained (post-hoc), and those where the explainability is part of the model architecture and design (intrinsic). Post-hoc methods include model-agnostic approaches such as LIME [49] and KernelSHAP [50], that obtain the explanations by perturbing the inputs and observing the changes caused on the outputs; also, model-specific methods such as Grad-CAM [51] and Deep SHAP [50], that use the parameters of neural network models to obtain the saliency maps. Among intrinsic methods,

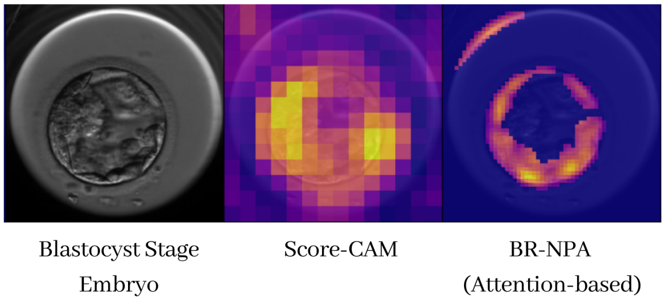

Figure 4.2: Saliency maps for Score-CAM and BR-NPA XAI methods from [8].

we can find attention-based CNNs, where the explainability is obtained from attention layers. A different approach is the use of Speeded Up Robust Features (SURF) [52] to extract local visual features, in conjunction with Gaussian Mixture Models (GMM) to obtain a Fisher vector per image [53].

Ref.	Task	XAI Method(s)	Clinically Assessed XAI
[32]	Quality Grading	Grad-CAM++	-
[54]	Quality Grading	Grad-CAM	-
[53]	Quality Grading	BVW, Grad-CAM	-
[55]	Embryo Segmentation	Grad-CAM	-
[56]	Development Stage Identification	Grad-CAM, SHAP, LIME	✓
[57]	Quality Grading	Grad-CAM	-
[8]	Development Stage Identification	B-CNN, Attention Branch Network, InterByParts, Grad-CAM++, RISE, Score-CAM, Ablation-CAM, AM	-
[34]	Fetal Heart Pregnancy, Live Birth	Grad-CAM, SHAP	-
[42]	Live Birth	Attention Branch Network	✓
[58]	Quality Grading	Grad-CAM	-
[59]	Quality Grading	CAM	-

Table 4.4: Papers on Explainable AI. *Clinically Assesed XAI* reports studies where experts evaluate the explanations provided by XAI techniques from a clinical perspective.

Only two studies have involved clinicians in the evaluation process. In the study conducted by Sharma et al. [56], three embryologists assess the biological relevance of heatmaps generated by Grad-CAM and LIME. The findings suggest that LIME explanations may be less consistent with biologically significant regions, while SHAP could potentially identify reasons for misclassification between adjacent cleavage stages. Sawada et al. [42] employ an Attention Branch Network for Live Birth Prediction, allowing experts to visualize relevant embryo features through the attention mechanism. These features are assessed by clinicians using saliency maps, but the evaluation reveals no common visual features associated with the predicted outcomes of live or non-live birth.

Some researchers utilize XAI techniques solely for visualization purposes without involving clinicians in the evaluation process. Paya et al. [57] employ Grad-CAM and note that the key regions highlighted by the model align with common interpretations made by clinicians. Diakiw et al. [32] utilize Grad-CAM++ for visualization while statistically correlating model outputs with the Gardner score. Thiramalaju et al. [59] directly use CAM saliency maps and observe that their model focuses on well-known features such as cellular fragmentation, blastomeres, or vacuoles. In a different approach, Kallipolitis et al. [53] compare Grad-CAM with SURF, finding that the former may erroneously focus on irrelevant areas according to the authors. However, these studies lack the involvement of clinicians in evaluating the XAI results.

Enatsu et al. [34] use Grad-CAM to detect morphological features that contribute to the classification, while incorporating SHAP to account for relevant metadata influencing the model's decisions. Their findings indicate that embryo images are the most effective predictor for fetal heart rate, followed by age and pregnancy history. Additionally, Arslan et al. [55] propose a multi-scaling architecture that segments embryos into distinct regions and applies Grad-CAM to different layers of the network, enabling visualization of feature attributions in various segmented parts of the embryo. This approach provides a higher level of granularity in the explanations by visualizing saliency maps at different segmented regions.

To the best of our knowledge, other very important challenges in this field, such as *Blastocyst Prediction* or *Ploidy Detection*, are not addressed in the context of explainability. This highlights the need for further research in these tasks.

Chapter 5

Discussion

So far, a literature review has been designed and conducted (Chapter 3) and the studies have been analyzed for a comprehensive understanding of the State of the Art (Chapter 4). In this chapter, the contributions are completed by analyzing and discussing the intricacies AI-assisted embryo selection. The discussed contributions are organized in three sections: IVF-related data for model development, model evaluation and ethical considerations of AI-Assisted embryo selection.

5.1 IVF-related Data for Model Development

In the field of IVF, the selection and quality of data are vital for the development and effectiveness of AI models. This section examines which data is commonly used (*i.e.*, images and metadata) and how it is integrated in the models.

5.1.1 Data Selection based on Intended Use

The selection of data for training models is strongly linked to their intended use in a real clinical context. For example, a model trained for the prediction of blastocyst formation should not be used to predict pregnancy or live birth. Ideally, experts should be involved in almost every step of the process. The correct definition of the intended use empowers users and enables the safe use of the system. Disregarding the intended use may lead to unpredictable, unreliable or unsafe system behavior and, as a consequence, a direct impact on the patient and their well-being. The intended use of a model can be defined from two different perspectives.

Embryo types and image selection

First, embryologists should define what **types of embryos** they want to screen, which should be reflected in the selection of images for model training. For example, for *Pregnancy Prediction* and *Live Birth Prediction* tasks a choice can be made between using images from a single day (D5), or using images of the complete development of the embryo up to day 5 (D0-D5). However, it is possible that experts may want to have the extra functionality to predict these events at D3, such situations should be considered in the design phase.

Another aspect related to intended use is the possible **pre-selection of embryos** by experts prior to the use of the model. Not all embryos are of good quality, some may present malformations, defects or suffer arrested development for unknown reasons. Defining whether an expert pre-selection of embryos will be done is vital to ensure that the capacity of the model is not exceeded by cases that have not been seen during training. Some studies in the literature aim to automate the embryo selection process entirely and do not rely on pre-selection by embryologists [39, 33, 46], while others focus on discriminating only between previously transferred embryos [38, 37, 36, 60, 30, 45, 46]. Systems that do not require prior review of embryos by experts should be used with caution, as *automation bias* may arise.

Clinical settings

Aspects related to the clinical setting should also be defined. For example, it has been seen that images of embryos that have been fertilized by IVF differ from those fertilized by ICSI, since in the former case, spermatozoa may remain visible in the well. Therefore, it should be defined whether, in the clinical context, images of embryos obtained by different **fertilization methods** will be used. Similarly, the method of embryo transfer should also be defined, whether the embryos are transferred **fresh** or transferred after **cryopreservation**. In addition, other aspects such as **genetic testing** must be specified in advance. Tests such as PGT-A have been shown to produce a morphological alteration to the embryo due to the invasiveness of the operation; this may create differences between images of embryos that have been tested PGT-A and embryos that have not.

Tables 3.1, 3.2, 3.3 and 3.4 show the data related to these variables that have been defined in the studies. At a glance, the general lack of information disclosure can be observed. This highlights the lack of robustness and reliability of the models that have been developed. Another consequence of not defining this information is that comparison between studies becomes practically impossible, since the characteristics of the data on which the models have been trained may differ greatly.

5.1.2 Pre-processing of Embryo Images

Embryo images obtained from microscopes can be susceptible to various forms of corruption, which requires careful consideration and pre-processing solutions during the model development process (Figure 5.1). As stated in the previous section, developers must explicitly state the intended use of their model to set expectations regarding its capability to handle corrupted and low quality images. It is important to clarify whether the model is designed to be capable of dealing with such images or if clinicians are expected to pre-select high quality images for the training of the models.

Image resolution

The visibility of biological structures and their changes during embryo development plays a crucial role in their assessment. However, the resolution of the images used for training the model can pose challenges to this visibility. In the literature, both low-resolution images (e.g., 50x50 px as seen in Miyagi et al. [43]) and higher-quality images (e.g., 480x640 px as utilized by Enatsu et al. [34]) have been employed. However, it is difficult to know the impact of resolution since, on the one hand, ablation studies with different resolutions have not been done in the same work, and on the other hand, comparison between studies is impossible due to the use of different data and models. On the one hand, the selection of lower quality images may lead to loss of information. On the other hand, utilizing high-quality images can enhance performance, but it may require extensive computational resources while increasing the dimensionality of inputs, with the consequent implications for the learning procedure (*e.g.*, curse of dimensionality, increase in model complexity, overfitting and generalization challenges). We recommend to work at the highest resolution possible if feasible with the aim of not losing key information.

Embryo detection in the well

Embryos are cultured in a well, which has a round shape and is visible in the images taken by the microscope (Figure 1.2). One common issue is the presence of impurities scattered around it; these impurities can result from the detachment of granulose cells from the zona pellucida, gel residue or sperm in the case of standard IVF procedure, leading to artifacts in the image. Furthermore, the position of the embryo within the culture well is not fixed, as it tends to move during the developmental process and thus, can be a source of noise.

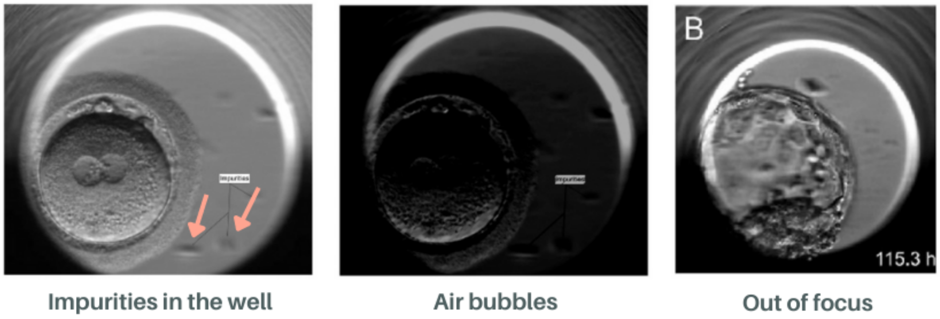

Figure 5.1: Examples of bad-quality images. From left to right: granulose cells impurities in the well, air bubbles that turn image dark and out of focus image.

These type of challenges call for solutions that allow the detection of the embryo in the well to isolate it from its environment. Some researchers opt for traditional computer vision techniques for embryo detection [44], while others employ DL-based alternatives. For instance, Wang et al. [58] use *YOLO v3*, a DL-based object detection algorithm which is used to detect and position the embryo at the center of the image. Kan-Tor et al. [30], on the other hand, train an image segmentation model, U-Net specifically, to isolate the embryo form the environment. While traditional computer vision techniques are less complex and have lower computational requirements, DL-based techniques can be more robust but in turn involve more complexity and higher computational demands. In future research, it is recommended to explore the effectiveness of newer techniques, such as Visual Transformers, which are known to be less vulnerable to changes in the embryo's position within the well. Additionally, the evaluation of novel segmentation models, like SAM [61], could prove valuable, as it has demonstrated high performance even with limited training examples.

Manual curation of images

Instead of developing models capable of handling image quality, some studies resort to manual curation of images. For example, Liao et al. [29] and Bormann et al. [36] discard misplaced embryos and images with occluded embryos, while Chavez-Badiola et al. [35] select images based on criteria such as sufficient light, sharp focus, absence of instruments or artifacts, and complete visualization of the embryo. Overall, manual curation varies in its approach to address issues like misplacement, occlusion, and image quality. Manual curation has nonetheless several limitations: it places an added workload on embryologists and it adds a implicit bias by this curation process, which makes systems brittle and unstable in the presence of noise. Therefore, it is believed that solutions such as those presented above are more appropriate and more robust.

5.1.3 Integration of Images and Metadata

Developing models that include both images and metadata can result in higher performance, particularly for long-term prediction tasks like predicting pregnancy or live birth, where many other factors influence the outcome. These factors include patient related data (*e.g.*, patient age, endometrial thickness, hormone levels), clinical data (*e.g.*, embryo culture conditions) or morphokinetical data (*e.g.*, cell symmetry, presence of vacuoles).

There are few cases in the literature where these parameters are included. Enatsu et al. [34] use both static images from blastocysts and clinical data for the development of their *Pregnancy Prediction* task model and report higher performance than simply using the embryo images. One approach to combine images and metadata in the models is to process them separately before integrating the extracted representations at a later stage. This can involve separate branches for image and metadata inputs, followed by fusion layers for combining the information. This approach has the benefit of increasing the transparency of the model, as the relevance of visual and non-visual features can be assessed separately. Furthermore, attention mechanisms could be employed to focus on different aspects of the image and metadata. For example, in an embryo image, the model should attend to specific regions of interest (*e.g.*, cytoplasm) based on the metadata information (*e.g.*, fragmentation rate). In any case, the performance of this type of model should be studied in the form ablation studies to evaluate the contribution of each of the modules.

5.1.4 Integration of Time Dimensionality in TL Videos

TL videos are able to capture kinetic information of embryo development, which allow for a more comprehensive understanding of evolution throughout different stages of embryo growth. However, developing models that utilize TL videos as input introduces greater complexity and requires the development of more advanced models. The predominant approaches observed in the literature include the utilization of 3D-CNNs, LSTM networks, or a combination of both. However, a comprehensive analysis of the individual properties and performance characteristics of these techniques is lacking, as no ablation studies have been

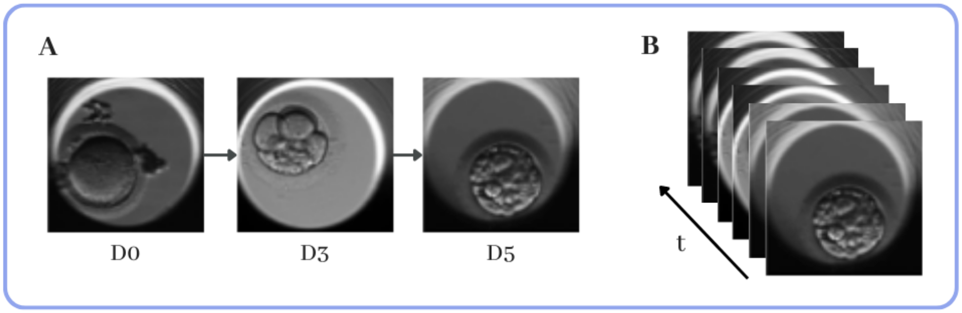

Figure 5.2: A: images of embryo development from D0 to D5. B: stacked frames of a TL video of an embryo.

conducted to date. Thus, determining the optimal technique remains an open question.

There are cases where even if developers have access to TL videos, they choose to process each frame as static images. This approach, while simplifying the problem, can lead to the loss of relevant information contained in the temporal dimension of the data, as well as to higher tendency towards overfitting. That is the case of the work of Sawada et al. [42], where they treated each frame of the TL videos as individual samples. Consequently, their dataset initially comprised of 470 unique embryos, is converted into a 141,444 instance dataset; this causes to be limited variability among the samples.

The relevance of the temporal factor for embryo assessment is well known, as its present in most embryo assessment guidelines. It is important to note that analyzing TL videos is not a trivial problem, as each video can consist of hundreds of frames, around 700 images in a 5 day development span of time with an image acquisition frequency of one image per 10 minutes. There are works such as [29, 30] which use the whole TL video for training the models. In turn, if only a subset of the TL video is selected, the sampling frequency used for the selection of images that will be used as input to the model has to be defined. Communication with the experts on this subject allows to define which phases or time intervals are important in the evolution of the embryo, which will allow the developers to select a optimal sampling frequency. The approaches to address the challenge of time dimensionality in TL videos can be divided into two main groups: development of specific models trained for detecting the most informative frames and selection of specific time frames in an arbitrary manner.

Chen et al. [27] propose a model that aims to identify key-frames of variable length within the video, which in turn captures relevant temporal information while reducing the overall dimensionality of the data. This model is composed of both a CNN and an LSTM, in addition to a policy that is responsible for evaluating the relevance of each temporal frame. It should be added that, in the case of using this type of model, the selection of frames should also be evaluated by experts during the design phase to validate the relevance of the frames from a clinical perspective. As an alternative, works by Lassen et al. and Berntsen et al. [46, 33] arbitrarily select the frames or the used sampling frequency. While this is a simpler approach, it has the limitation of selecting low informative frames if the sample rate is set too low.

Using images of TL videos as input to the network is not the only way to process temporal information. Silva-Rodriguez et al. [44] extract morphokinetic parameters from the image sequences instead of using the TL videos directly for model training. A limitation of not using the image directly as a predictor and, instead, using morphokinetic parameters extracted from the image, is that these parameters have to be sufficiently representative of the images for the model to make accurate predictions.

In summary, it is advisable to avoid treating individual frames of TL videos as separate samples to mitigate issues like overfitting. Instead, it is recommended to develop more complex architectures that can process the entire video of a specific embryo as a single sample. Moreover, when sampling video frames, it is strongly recommended to validate the selection in consultation with clinical experts. This collaborative approach ensures the inclusion of relevant frames and enhances the overall accuracy and reliability of the models.

5.1.5 Integration of Multiple Focal Planes

The integration of multiple focal planes presents another layer of complexity in the development of embryo selection models. TLIs have the capability to capture images from up to nine different focal planes, depending on the specific brand and equipment used. Each focal plane can be understood as a different slice of the embryo, where different structures can be seen (Figure 5.3).

Clinicians typically rely on the equatorial focal plane to assess embryo development, as it provides a generic view for evaluating the embryo's appearance. However, different focal planes can offer additional information, particularly in cases where cells overlap, making tasks such as cell counting more challenging. By incorporating multiple focal planes, it is possible to capture a more compre-

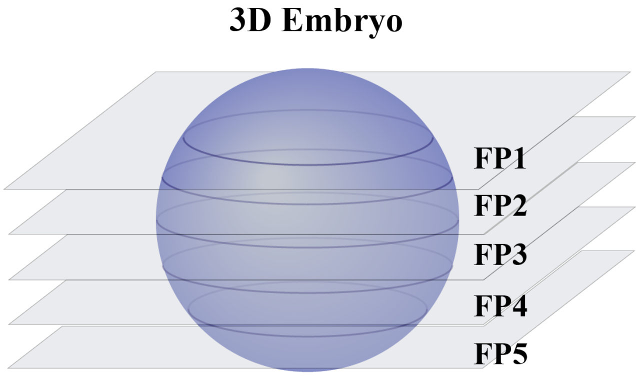

Figure 5.3: Five hypothetical focal planes of an embryo.

hensive understanding of the embryo's characteristics and potentially improve the performance of the model.

The simplest and most frequent approach found in the literature review is the arbitrary selection of a single focal plane. For example, Lassen et al. [46] and Bertsen et al. [33] select a single central focal plane of each embryo, even though having access to the data from several focal planes. Due to the large amount of data with which these models have been trained, it would have been interesting to see a more complex architecture capable of processing different planes to see if there is an added value in terms of performance.

Nonetheless, two studies has been found that processes multiple focal planes, namely Wang et al. and Kan-Tor et al. [58, 30]. Wang et al. [58] propose three different methods to deal with multi-focal images: (1) a shared network extracts features from the 11 focal planes, which are stacked together and used for final classification; (2) predictions for each focal plane are made independently and a voting mechanism is used for the final classification; and (3) the model calculates the overall sharpness of every focal plane and chooses the top 3 clearest images, which are then fed as different channels to the network. This last model is the one which shows best performance and is used for *Blastocyst Formation Prediction* task, although it could also be adapted for *Pregnancy Prediction* and *Live Birth Prediction* tasks.

The previous methods methods are capable of processing one or more focal planes of the same embryo, although none of them deal jointly with temporal dimensionality, all of them process static images. Only one study has been found that integrates both types of data. Xie et al. [28] develop a custom architecture which first extracts features from different focal planes in a single frame, and combines them sequentially with features from future frames. An ablation study, shows that performance is highest when the model combines both temporal and focal plane information. Since the benefit of integrating both temporal and focal information has been observed, training these models for other tasks could also represent interesting lines of research.

5.1.6 Impact of Class Imbalance

Class imbalance in IVF datasets has significant implications for the development and evaluation of AI models. In the context of traditional IVF, the success rate of transferred embryos is typically around 30% [7]. Consequently, datasets which contain data of transferred embryos tend to have several times more negative outcomes than positives.

To properly assess the impact of class imbalance, it is crucial to report the test set prevalence when reporting results, which refers to the proportion of positive samples in the dataset. This allows for evaluating class imbalance and comparing model performance against random chance or naive guessing. For example, in a dataset with a prevalence of 30%, naive guessing by always predicting "negative" would yield a naive accuracy of 70%. Therefore, model performance should be compared to a random chance of 70% instead of the typical 50% [10]. The class balance and distribution of images/videos in the current literature are presented in the *Labels Distribution* column of tables 3.1, 3.2, 3.3, and 3.4. However, it is worth noting that not all papers include this information, making it challenging to assess their performance and correct selection of evaluation metrics.

Addressing class imbalance requires careful consideration. Undersampling the negative outcomes is a simplistic approach that comes at the cost of losing valuable data, which is already limited in the field of embryo selection; thus, this solution should be disregarded. Alternative strategies include oversampling the positive outcomes, as Huang et al. [41] do in their study. Weighted sampling techniques and adjusting the optimization algorithm to assign equal importance to misclassifying positive and negative examples [10] are other viable options more suitable for this problem. Other novel techniques such as synthetic data generation which resemble the original data could be explored in the future to account for data imbalance.

On another note, several studies have been identified where DL models are trained using unbalanced datasets [42, 39]. For instance, Tran et al. [39] utilize a heavily unbalanced dataset, with negative pregnancy instances comprising

92% of the entire data. However, training a DL model for embryo selection using such unbalanced data presents significant risks. The model may exhibit bias towards the negative class, resulting in skewed performance and lower accuracy for the positive class. Additionally, the model's ability to generalize becomes limited, as it struggles to predict outcomes for the underrepresented class.

In summary, it is crucial to avoid training models with unbalanced data, and the techniques employed to address this issue should be both robust and capable of mitigating the imbalance without sacrificing valuable information.

5.2 Model Evaluation

The evaluation of AI models in the context of embryo selection plays a crucial role in assessing their effectiveness, reliability, and generalizability. In this section, various aspects of model evaluation are analyzed, ranging from the selection of appropriate evaluation metrics to data splitting strategies. Furthermore, the importance of evaluating models beyond their performance on test sets is discussed in order to gain a deeper understanding of their real world applicability.

5.2.1 Evaluation Metrics for Embryo Selection

Several evaluation metrics are employed to assess the performance of AI models in embryo selection. Pregnancy prediction will be used as an example in order to understand the meaning an impact of the different evaluation metrics. In this context, a false positive translates to a failed implantation or miscarriage after transfer of a chosen embryo, whereas a false negative translates to a missed pregnancy because the embryo was incorrectly deprioritized for transfer.

Accuracy, which measures the proportion of correct implantation predictions, is commonly used as a performance measure as it is easy to understand. However, it should be carefully used when the dataset being evaluated is unbalanced. For example, Chen et al. [62] use a highly unbalanced dataset and report an accuracy of 91%, which can be highly misleading for the reader. Hence, it is essential to compare reported accuracies to a baseline that reflects naive classification performance [10].

Positive predictive value (PPV), also precision, and **negative predictive value (NPV)** are metrics that describe the proportion of positive predictions that were in fact pregnancies and the proportion of negative predictions that were in fact failed implantations or miscarriages. **Sensitivity**, also recall or True Positive Rate (TPR), and **specificity**, also True Negative Rate (TNR), describe the proportion of positive pregnancies that were predicted correctly as positive and the proportion of negative pregnancies that were predicted correctly as negative, respectively. True negatives represent non-viable embryos that are important to consider in order to minimize the time to pregnancy and associated costs.

While many prediction models generate continuous predictions, binary values are often obtained through dichotomization, which involves setting a threshold to classify predictions into binary outcomes (*e.g.*, pregnancy/no-pregnancy). However, this dichotomization process may discard valuable information and assumes a single clinically relevant threshold [63]. Therefore, other metrics related to discrimination operate on continuous prediction values. The **area under the curve (AUC)** of the **receiver operating characteristic (ROC)** is a metric that summarizes the model's performance across the entire range of scores, independent of a specific threshold. Almost all studies evaluate their models based on this metric.

Few metrics have been reported to rank embryos within a single patient cohort. Chavez et al. propose [35] the **normalized discounted cumulative gain (nDCG)**, which measures the ranking quality within a cohort by considering the relevance and position of embryos in the sorted list of model scores. In the study, relevance is determined based on the outcomes of preimplantation genetic testing (PGT). Diakiw et al. [32] also provide a metric to reflect the performance in embryo ranking, **Time to Pregnancy (TTP)**. TTP is calculated as the position of the top viable embryo in the ranked cohort.

It must be remembered tha there are expenses associated with misclassification in embryo selection, and the optimal compromise between these may differ between clinics and patients, as the trade-off is defined by multiple factors. In IVF, these could involve financial costs related to embryo cryopreservation and emotional costs of patients related to transferring embryos that most likely will not result in pregnancies or financial costs. [10].

5.2.2 Data Split Strategy

An important decision related on the evaluation of the model is the selection of the data split strategy. Typically, a dataset is divided into train, validation and test splits, but special attention must be placed when working with medical data. The dataset should be divided at the patient or treatment level to avoid splitting embryos from the same patient into different subsets. Splitting solely

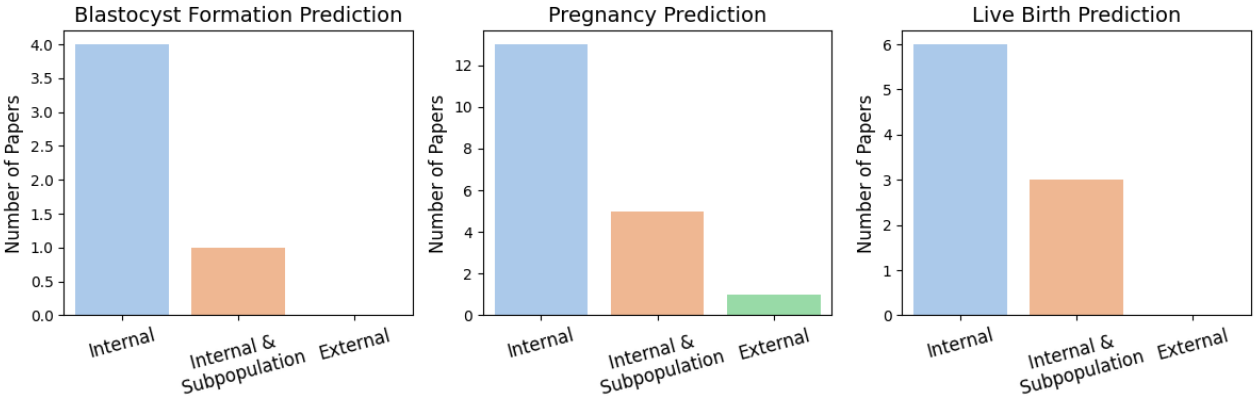

Figure 5.4: Validation type distribution across tasks in the literature.

at the embryo level may introduce bias due to the correlation between embryo images/videos and their associated outcomes [10]. Furthermore, for an even stronger data split strategy, the dataset could be split by time, training the model on an early time period and evaluating it on a later time period [63]. The way in which the model is validated is another crucial factor that demonstrates the model's robustness and generalization ability.

Internal Validation

The weakest kind of evaluation is the internal evaluation, where the test set is a data subset representing the same population and distribution as the training and validation set of the model. Using data from the same distribution does not allow to test how the model generalizes to different subpopulations in the data and also to account for possible biases. Most papers found in the review process fall into this group [27, 41, 34, 44], among others.

Internal Validation with Subpopulation Analysis

In cases where external data with different distributions is unavailable, researchers can still perform subpopulation analysis within the data split strategy. Subgroups based on factors such as patient age, body mass index, fertilization method, transfer protocol or ploidy status can be analyzed to detect potential biases. Even factors like different culture mediums are worth studying, because they have shown to impact on the model's predictions [18]. This approach allows for analyzing and detecting potential biases or the presence of confounders.

This approach is commonly used to assess the model's generalization capabilities to new healthcare facilities, such as different clinics or hospitals. Studies that collect data from multiple centers conduct this type of analysis. For example, studies by Tran et al. and Berntsen et al. [39, 33] performe cross-validation on different subgroups, including clinics, maternal age, insemination method, length of incubation, and fresh vs. cryopreserved embryo transfer. They report varying performance results across subgroups, indicating potential differences in generalization and biases within the dataset.

One notable study by Lassen et al. [46] comprehensively analyzes subpopulations in terms of age, insemination method, transfer protocol, year of treatment, and IVF clinic. Significant differences in performance are observed among age groups, transfer protocols, specific years of treatment, and certain clinics. These results underscore the importance of performing this type of validation in all the models developed, since failure to do so could discriminate against or disadvantage certain groups of patients.

External Validation

Moreover, a more rigorous evaluation method known as external validation assesses the model's performance in a completely different setting, such as a new clinic, time period, country, or population that was not part of the model's development [63]. For a pure external validation, the new setting, like a clinic, must be completely separate and independent during the model's development, as exemplified by studies employing "double-blind test sets". That is the case of Ver-Mileya et al. [38], where they test the generalization capability of their model to independent clinics that did not provide any data for training. Results show the model accuracy to be marginally lower due to the introduction of inter-clinic variability, which may have affected efficacy due to varying patient demographics and different equipment and methods for image acquisition.

Figure 5.4 illustrates the distribution of validation methodologies employed in the literature across different tasks. The data highlights that a significant portion of studies primarily rely on internal validation, while only a subset of these papers conduct subpopulation analysis. Furthermore, only one study goes a step further and performs an external validation of their model. This data emphasizes the need for stronger evaluation methodologies in future research.

5.2.3 Evaluation of Predictive Performance over Time

The evaluation of predictive performance over time is a critical aspect of assessing the robustness and reliability of AI models used in embryo selection. When a model utilizes TL videos, where the images capture different developmental stages of embryos, it becomes crucial to understand how the model's predictions may vary based on the specific temporal frames.

For instance, if a model has been trained on a dataset that includes images ranging from the cleavage stage (D3) to the blastocyst stage (D5), it is necessary to evaluate its performance when using images from different time intervals within that overall time span. This evaluation helps determine if the model's predictions remain consistent across various stages of embryo development. By systematically testing the model's performance using different time intervals, researchers can gain insights into its temporal sensitivity and assess its suitability for clinical use.

Lassen et al. [46] asses the predictive performance over time by evaluating the model at 12 hour intervals. Results show a significant improvement in predictive performance for later predictions on D2 and D5. This correlates with the decision of clinicians of selecting the embryo at later stages, where embryos appearance is more informative. Other works such as Kan-Tor et al. [30] also perform this kind of study, and they show that for *Blastocyst Formation Prediction* task AUC increased monotonically with time of prediction.

The need for evaluating predictive performance over time becomes particularly important to address potential misuse of the system in a clinical setting. Clinicians may unintentionally introduce images from only the beginning or the end of the developmental sequence, deviating from the intended usage of the model. By conducting evaluations that simulate such scenarios, researchers can identify any performance limitations of the model when used outside the expected temporal range.

5.2.4 Model Evaluation inside a Clinical Context

The evaluation of AI models often concludes upon achieving satisfactory performance on a test set. However, this limited evaluation raises concerns regarding the clinical validity of these models. It is crucial to assess their usefulness and practicality from the perspective of clinicians, as their acceptance and integration into clinical workflows greatly influence their real-world impact. Therefore, comprehensive evaluations should involve testing the performance enhancement of these models when utilized as DSSs by clinicians.

Although one of the primary motivations behind the development of AI models is to ensure their robustness in handling inter- and intra-observer variability, it is expected that the performance of these models is not lower than experts'. However, to accurately assess the performance offered by AI models, it is crucial to compare their results with the current methodologies employed by clinicians.

Nonetheless, only a limited number of studies in the existing literature conduct such evaluations, where the model's performance is directly compared to a group of experts using traditional evaluation methods on the same test set.

This kind of evaluation can lead to unexpected results, as demonstrated by Sawada et al. [42]. They compare the performance of the DL model and the performance of embryologist using traditional Gardner score. They report that the two evaluations are not concordant, suggesting that the AI model may focus on embryo features other than the ones assessed by embryologists. Nonetheless, they show that using the DL model as a DSS, in conjunction with conventional morphological evaluation, leads to improve performance for *Live Birth Prediction* compared to selecting embryos solely based on the AI system or conventional morphological evaluation alone.

Conversely, Bornmann et al. [36] demonstrate that their DL model can outperform 15 different embryologists in identifying embryos with implantation potential. Similarly, Wu et al [54], test the clinical value of their embryo quality grading model by asking 5 embryologists to grade the embryos in a test set by the aid of the model. They show that embryologists improved the AUCs by between 2% and 7%, thus also confirming how these systems can be helpful in a clinical setting. These studies further emphasize the clinical value and potential of these AI models.

In conclusion, evaluating AI models beyond their performance on a test set is crucial for assessing their clinical validity. While the aim of these models is to enhance the decisions of experts when used as DSSs, only a limited number of studies perform this kind of analysis. Further research and comprehensive evaluations are needed to fully understand their benefits and limitations.

5.3 Ethical Perspective of AI-Assisted Embryo Selection

The use of AI-Assisted embryo selection systems has both medical potential and ethical concerns. The integration of these models in human reproduction introduces new possibilities and complexities, giving rise to ethical questions for IVF patients, clinicians (especially embryologists), and society as a whole. In this section, a study of the different ethical dimensions that are involved in this complex topic is carried out. To this end, the framework proposed by [25] is taken as a basis, which combines two sets of principles: human-centric AI ethics and patient-centric principles of bioethics. Contributions from [24, 26] have also been insightful for the development of this section.

5.3.1 Transparency and Patient Autonomy

The principle of procreative or reproductive autonomy refers to the freedom of prospective parents to decide when, how, and under what circumstances they want to have children [25]. Transparency plays a vital role in the ethical use of AI in assisted reproduction. Prospective parents have the right to know the technical details of how their child is conceived through AI models. According to the Ethics guidelines for trustworthy AI developed by the European Commission's High-Level Expert Group on AI [64], individuals should be aware of interacting with an AI system and informed about its capabilities and limitations.

However, complete technical transparency may overwhelming and difficult to understand by patients. Instead, prospective parents should receive appropriate information from the fertility clinic or clinician about the functionalities of the AI models for embryo selection, the roles of AI and embryologists in the procedure and the specific benefits, risks, and limitations involved [25]. This empowers patients to make informed decisions and maintains a collaborative relationship between clinicians and individuals seeking reproductive assistance. By providing patients with clear and easy to understand explanations of certain decisions, they can be more involved and better equipped for the decision-making process regarding their own care.

Furthermore, an informed consent is vital in this process. According to [25], an appropriate informed consent requires that the IVF clinician appropriately inform prospective parents about: (a) embryo evaluation and selection to be performed automatically using an AI tool; (b) the distinguished, respective roles of AI and the embryologist in the process (*e.g.*, decision-support-system, automated system); and (c) specific utilities as well as potential risks and limitations.

5.3.2 Trust and Explainability

Explainability of AI systems is fundamental, specially in this sensitive domain, and a lack of it can pose a significant challenge and undermine trust among clinicians. The opaque nature of these models makes it difficult to understand how they arrive at their decisions, which in turn hinders the ability to explain and justify their recommendations. This lack of transparency can create skepticism and reluctance to fully trust AI systems in clinical decision-making processes.

Furthermore, the successful integration of AI-assisted embryo selection systems in the medical context heavily depends on the acceptance by the clinicians and, consequently, their trust in them. According to [65], there are several levels of trust which fall along a spectrum, ranging from complete distrust to over-reliance on AI systems. Studies have shown that over-reliance on the suggestions can cause clinicians to take less initiative [66] and also to be more likely to accept incorrect diagnosis [67]; this is known as *automation bias* [68]. Moreover, heavy reliance on AI tools might lead to the deskilling of embryologists [25]. On the other side of the spectrum lie clinicians which do not trust an algorithm that they do not understand [69], a phenomenon known as *algorithmic aversion*.

In the current literature, a limited amount of research comprehends the use of AI-assisted IVF in combination with XAI methods; this studies have been presented in section 4.2. In many cases XAI is only considered through the illustration and minor discussion of a few saliency maps [55, 53, 59, 54]. This approach is not without flaws. It can induce *confirmation bias* (*i.e.*, a clinician may only review the evidence that supports its own hypothesis). Others take one more step and analyze the map activations in order to correlate them to morphological features [56, 58] or to the objective [34]. Few share the saliency maps with expert embryologists for evaluation [56, 42]. Meanwhile, several studies conclude that saliency maps should not be used as the sole source of explainability in high risk medical domains [70, 71].

In summary, the explainability of AI systems in the field of embryo selection is crucial for building trust among clinicians. Sharing XAI results with expert embryologists and contextualizing their interpretation under clinician supervision is essential for meaningful and reliable explainability in this sensitive medical domain. Furthermore, the study of explainability within topic has served as a basis for the development of a publication submitted to *CCIA* conference, to which few members of the HPAI research group, including the author of this project, have contributed.

5.3.3 Non-Maleficence: from the Lab to the Clinic

The principle of non-maleficence is highly relevant in the context of AI-Assisted embryo selection systems. This principle emphasizes the importance of taking precautionary measures during the research phase to ensure the well-being of participants, including the potential children conceived through assisted reproductive technologies [25]. The application of non-maleficence requires that the benefits of these technologies outweigh the associated risks, maintaining a reasonable risk-benefit ratio.

One of the main limitations of current AI-assisted embryo selection is the lack of Randomized Controlled Trials. As [24, 25, 26] point out, the field of reproductive medicine largely relies on small retrospective studies, lacking RTCs to validate and optimize the utilization of AI models. Introducing AI-assisted embryo selection into clinical practice poses its own challenges, requiring successful and ethically approved RCTs. While one trial has been registered, it is important to wait for trial results before implementing the technology. RCTs are crucial for evaluating new interventions and ensuring patient safety. Thus, until RTCs show the true potential of this systems, the real validity will remain unknown.

Furthermore, the focus of research should be on maintaining the autonomy of prospective parents in the experimental group, ensuring transparency in the informed consent process [25]. In later stages of the research, close monitoring of pregnant women and child follow-up should be conducted to identify any potential long-term effects for both the mother and the conceived children.

5.3.4 Beneficence for Patients, Clinicians and Society

AI-Assisted embryo selection models holds promise for potential benefits such as improved pregnancy success rates or reduction of healthcare costs. This offers valuable advantages to various stakeholders, including prospective parents, particularly women undergoing IVF, as well as embryologists, fertility doctors, and society at large. These AI systems can greatly benefit the physician-patient relationship by allowing physicians to dedicate more time to strengthen their interactions with patients [26]. Additionally, this technology holds potential benefits specifically for the patients. The increased efficiency of treatment enables better life planning, including personal and professional trajectories, granting women greater control over their reproductive journey [25].

Furthermore, these systems can have great impact at a societal level. The integration of AI with embryo selection raises concerns regarding its association with eugenics. This topic is addressed in the "Beneficence" section because some could argue that the selection of favorable human traits can be seen as beneficial for society, although this viewpoint is highly debatable. AI systems offer prospective parents the opportunity to select the most viable embryos, effectively allowing them to choose the "best" offspring in terms of viability. While the current focus is primarily on selecting embryos based on viability, it is important not to disregard the potential for more elective and nuanced selection of genetic traits, which directly relates to the concept of eugenics [25].

This approach could be seen as discriminatory against individuals with disabilities and convey a negative message about the value of their lives. For instance, screening for conditions like Down Syndrome has been criticized for expressing a negative view about the worth of people with Down Syndrome. While this objection applies not only to AI selection but also to clinical selection in general, AI has the potential to significantly expand the scope of this concern [24].

5.3.5 Responsability and Accountability

The use of opaque AI models raises ethical and legal accountability concerns [24], particularly when clinicians can not explain the decision-making process. This creates a "responsibility gap", and without established accountability mechanisms it is challenging to determine who should be held responsible for any potential harm. For example, in cases of sub-optimal embryo selection or injury due to model recommendations, the decision-making process must be explainable to patients seeking to understand what happened or, in more extreme cases, seeking compensation for the damage caused. As of today, distrust in AI applications in medicine also comes from doctors fear of legal repercussions if something goes wrong due to unclear liability regimes [65]. If clinicians base their decision on these opaque AI models, the evaluation of the decision-making process and, consequently, the determination of who is responsible will be greatly hampered.

5.3.6 Reproductive Access and Social Justice

Works such as [26, 25] have studied how social justice in terms of eligibility and access is addressed in the topic of AI-Assisted IVF. The affordability of AI in IVF and its impact on accessibility are key factors in analyzing social justice. The cost of AI in IVF should not become a barrier to treatment access, particularly for individuals with limited financial resources. While traditional IVF treatments are already unaffordable for many, the initial stages of implementing AI in IVF may further exacerbate this issue. However, one advantage of the implementation of this technology is the potential long-term reduction in treatment costs, leading to increased affordability and equal accessibility.

Nevertheless, to achieve cost reduction for patients and the healthcare system, the use of AI-based technologies in reproductive medicine must be efficient and not entail disproportionately higher costs. Tied to this idea, [24] puts forward the following problem: if a clinic decides to adopt a specific AI model for reproductive medicine, they would need to adhere to the ecosystem and protocols associated with that model, including ovarian stimulation regimens, the use of specific incubators, culture medium, and other variables. This effectively gives AI companies significant economic power over clinics, potentially leading to increased treatment costs.

5.3.7 Algorithmic Bias and Fairness

Throughout this study, various forms of bias have been identified and discussed, highlighting their presence in different contexts. One aspect of bias stems from a poor definition of the intended use of the model. Where image pre-processing techniques or the inclusion criteria of certain types of embryos is not clearly defined. Furthermore, factors such as genetic abnormalities, developmental stages, fertilization methods, and the use of fresh or frozen embryos can introduce unbalanced data within sub-populations, leading to biases against specific classes or traits. It is important to explicitly address these biases and consider sub-cohort analysis when developing AI models for embryo selection.

In addition to data-related biases, biases can also emerge during the implementation of AI systems in real clinical settings. Automation bias, for instance, refers to the tendency to unquestioningly accept incorrect diagnoses or recommendations generated by AI models. Confirmation bias is another concern, where clinicians may interpret medical results in a way that aligns with their preconceived notions or hypotheses. In summary, biases permeate various stages of the development and implementation of AI-assisted embryo selection systems.

These biases underscore the need for rigorous scrutiny and ongoing evaluation of AI models in reproductive medicine. Overcoming biases in AI-assisted embryo selection is no easy task, as it requires a comprehensive approach that encompasses data collection, model development, implementation, and evaluation, as well as compliance with ethical and legal standards through all the stages. Transparency and interpretability in AI models also play a significant role, as it has been discussed. Building models that provide insights into the decision-making process enables clinicians to evaluate and potentially correct biases. Addressing and mitigating biases is essential to ensure fair outcomes for patients and to maintain the trust and reliability of these AI systems.

Chapter 6

Conclusions

AI-assisted embryo selection is a highly promising field of research, yet it also presents multiple complexities. These have been overlooked in previous studies, raising doubts about their reliability. This project has aimed to bridge the gap in the current literature by analyzing both technical and ethical aspects that need to be considered in the development of these systems.

Special emphasis has been placed on the importance of involving expert embryologists throughout the model development and evaluation, ensuring that they are reliable and provide added value to their decisions. It has been learned that a lack of collaboration may lead to overlooking crucial aspects that could impact future patients. For instance, an important aspect of explainability is the subsequent validation of produced explanations by domain experts. There have been cases where generated saliency maps have failed to provide any value in terms understanding the model or enhancing the decision-making process. This highlights the need for an interdisciplinary approach in the matter.

Ethical aspects associated with AI-assisted embryo selection are also a topic of great relevance, and they should be considered for the development of such systems. In this work, we have discussed various ethical perspectives, from biases present in model development to transparency, accountability or potential societal impact. By incorporating these ethical dimensions, we aim to ensure fair and safe outcomes.

As a result of the analysis and discussion of the aforementioned aspects, this work proposes a series of recommendations for the development of AI-assisted embryo selection models:

1. Collaborate with clinical experts and end users throughout model development to incorporate their expertise and needs.
2. Define the task of interest and intended use of the model early on, specifying the types of embryos that will be screened and the clinical settings.
3. Prioritize models trained on pre-selected embryos to maximize performance and prevent automation bias.
4. From the start of the project, purposefully design the model with a focus on integrating explainability methods.
5. Build a robust pipeline which itegrates image pre-processing techniques and develop models capable of integrating information related to time and focal planes.
6. Report results using relevant evaluation metrics and integrate experts in the evaluation of the model *e.g.*, simulating the use of the model as a DSS.
7. Report data-related and model-related information using tools such as Datasheets for Datasets [72] or Model Cards [73] in order to increase transparency, and follow ethical standards such as the Assessment List for Trustworthy AI [64].

By following these recommendations, researchers and practitioners can work towards developing reliable and trustworthy AI models for embryo selection, ultimately improving clinical outcomes and reproductive healthcare practices.

Chapter 7

Future Work

The work done within this thesis is part of an interdisciplinary collaboration between the HPAI research group at the BSC and the Clinic Hospital in Barcelona. The initial objectives of this collaboration encompass the work presented in the preceding sections, which includes a comprehensive review and critique of all aspects related to AI-Assisted embryo selection. Additionally, the long-term goals of this collaboration involve the development of AI models for embryo selection, which will be heavily influenced by the findings of this thesis. The focus of these models comprise three key tasks that have been presented in the current study: *Blastocyst Formation Prediction*, *Pregnancy Prediction* and *Live Birth Prediction*.

7.1 Data Management

In the development of the AI models, both patient metadata and clinical data, along with embryo images, will be used. To ensure compliance with GDPR regulations regarding personal data protection, extensive documentation has been developed. This documentation specifies technical measures for access control and data storage. The responsibility for data anonymization lies with IDIBAPS, who performs this process at the hospital. Patient identifiers and any identifying metadata are removed to ensure the data's anonymity. The data is stored in the EmbryoScope, located at the same center, and managed by IDIBAPS.

The process of accessing the EmbryoScope data has been carefully outlined. Access to the device is restricted by the IDIBAPS security network. To enable BSC's access, IDIBAPS provides a certificate that allows secure and auditable data retrieval through the REST API of the EmbryoScope. This connection is established via VPN through the private network of the EmbryoScope. BSC securely stores the encrypted data obtained from the EmbryoScope on a dedicated storage device within their premises. Importantly, the data is never stored on personal computers, and it remains encrypted throughout the process, ensuring data security and privacy.

7.2 Use-case Definition

The first task to be tackled in this project is *Blastocyst Formation Prediction*. This task consists of predicting which of the embryos developed up to day 3 have the potential to develop to blastocyst stage at day 5. It is of great relevance since, in the case of patients with few eggs available or with a high probability that they will not develop until day 5, the transfer can be performed on day 3 when more embryos are available. In this section, a methodology for the development of predictive models for *Blastocyst Formation Prediction* is outlined.

Definition of Indented Use with Clinicians

The initial step in addressing this task involves establishing the intended use of the model in collaboration with experts from Clinic Hospital. The following aspects have been defined to guide the development process:

1. Consideration of clinical settings: The dataset encompasses embryos resulting from both ICSI and IVF. Additionally, the presence of fresh and cryptopreserved implanted embryos will be analyzed in addition to those that have been genetically tested for aneuploidy (see Section 5.1.1).

2. Clinician's pre-selection of embryos: It is crucial to determine whether the model will evaluate all embryo types, including both high and low quality, or if a pre-selection by clinicians is anticipated. In this specific use case, clinicians will perform a pre-selection of good quality embryos, enabling the model to discriminate among them (see Section 5.1.1).

3. Selection of images based on embryo state: The model should use TL images capturing embryo development from D0 to D3, enabling the incorporation of both morphological and kinetic features. Various approaches for integrating temporal information will be explored during model devel-

opment (see Section 5.2).

In the discussions with the experts, an essential aspect that has been emphasized is the need for explainability in this project (see Section 5.3.2). It has been established that the developed models should provide explainability through saliency maps for images and other feature attribution techniques for metadata. These explanations will undergo evaluation by the experts during the development phase to ensure that they highlight relevant features and facilitate their understanding of the model's functioning and decision-making process.

Data Pre-processing Requirements

The following data pre-processing steps have been defined for this project:

1. Data quality: Images of embryos obtained through microscope are subject to corruption and artifacts. Therefore, this project will explore methods to deal with these problems. For example, techniques such as segmentation can be explored to isolate the embryo from the well, as it may contain remains of the clinical process that can introduce noise into the model. Additionally, by isolating the embryo from the medium and cropping the image based on its contour, the issue of embryo movement within the well will be avoided, which can also introduce harmful variability to the model. Furthermore, other methods should be analyzed to address the challenges posed by blurry or darkened images (see Section 5.1.2).

2. Sampling frequency: Another important step of data pre-processing is the selection of the sampling frequency when dealing with TL videos. The minimum sampling frequency should be the one that allows to capture relevant morphokinetical changes in the embryo development. High sampling frequencies may result in the oversight of important indicators of inadequate development, such as rapid cell division (direct cleavage). Hence, sampling frequency will be discussed with Clinic experts (see Section 5.2).

3. Focal planes: Additionally, the selection of focal planes used for model training needs careful consideration. Clinic experts primarily evaluate images using the equatorial plane but may resort to other focal planes if the equatorial view is insufficient. Therefore, different sampling frequencies and the integration of various focal planes will be explored in this project (see Section 5.3).

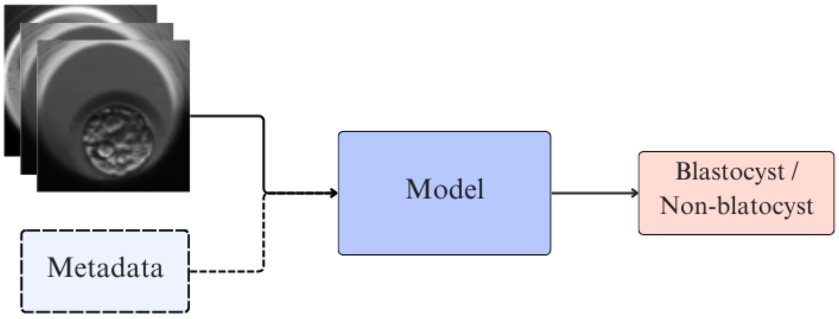

Figure 7.1: Illustration of raw image-based model. Dotted line indicates metadata can also be used as input.

7.3 Model Design

Due to the diverse range of data types collected in the process of embryo selection (time, focal planes, metadata), there is no single definitive model to address the problem. For this project, specifically focusing on the task of *Blastocyst Formation Prediction*, several models of varying complexity have been devised, taking into consideration the requirement of explainability for each of them.

Image-based and Image- and Metadata-based Models

The simplest yet less explainable model involves training a DNN directly on embryo images from D0 to D3 to classify them based on their potential to develop into a blastocyst (Figure 7.1). To handle the temporal dimensionality, the simplest option is to treat each frame as an individual sample and then aggregate the output of each frame, for example, using a voting system to obtain an embryo-level classification. Taking it further, the entire TL video can be used as input, and more complex methods like 3D convolutions can be employed to capture temporal information, directly yielding an embryo-level classification. However, this approach offers low explainability as saliency maps obtained through methods like Grad-CAM attribute the score to the entire image, often resulting in diffuse and challenging-to-interpret areas of high attribution.

An extended option for this model is to incorporate additional metadata collected during the patient's clinical process, such as age, biological indicators, or expert annotations on the embryo's appearance, alongside the images. This additional data could potentially improve the model's performance. It's important

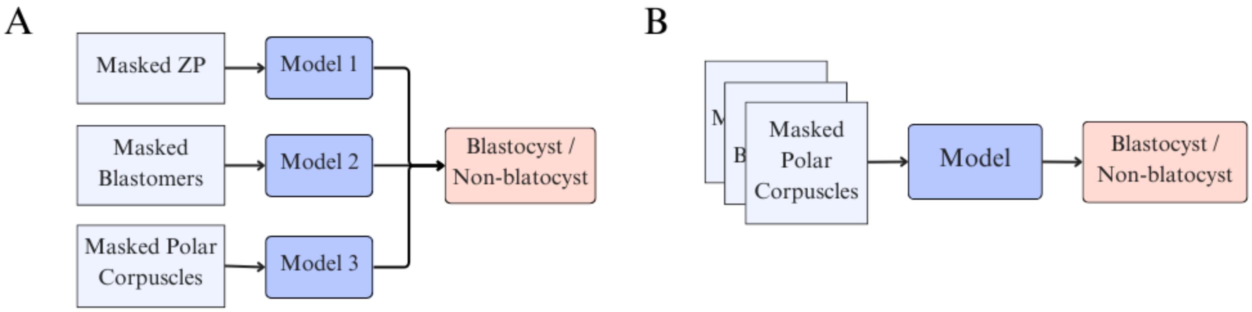

Figure 7.2: Segmentation-based models. A: different sub-networks are trained for each segmented image. B: segmented images are inputed to the same network as different channels.

to note that this model is designed to be simple and serves as a starting point for further development. Its purpose is to establish a baseline and pave the way for more advanced models in the future.

Segmentation Masks-based Models

An alternative that offers greater traceability of the final classification involves training specific sub-networks for different parts of the embryo. This approach includes segmenting the embryo based on its main morphological structures (*e.g.*, ZP, blastomeres, polar corpuscles) and using the resulting segmented images as input for these specialized subnetworks (Figure 7.2). By doing so, specific feature maps can be obtained for each relevant zone, enabling the creation of more specialized and potentially interpretable saliency maps. One drawback of this approach is that since the network only attends to isolated parts of the embryo, information regarding the collective appearance of the parts is lost. It must be added that metadata can also be used as input as in the previous example.

Therefore, an alternative to this network is to use the segmented parts of the embryo as different channels in the input. This means training a single model instead of separate models for each mask, as in the previous case. Additionally, this approach has the advantage that all parts of the embryo influence the final decision, while maintaining explainability as feature maps can be visualized for each channel, thus obtaining localized feature attributions for each structure of the embryo.

Model Evaluation

Finally, once the models have been trained, they will be evaluated to ensure their reliability and robustness. As discussed in the evaluation section, sub-population studies will be conducted to detect possible biases and confounders in the models. Sub-populations such as patient age or fertilization and transfer methods will be examined for this purpose. Additionally, since these models process TL videos, their predictive performance over time will be studied to assess their discrimination ability at different stages of embryo development.

Furthermore, to increase transparency, tools such as Datasheets for Datasets [72] and Model Cards [73] will be employed. Datasheets for Datasets provide detailed information about the dataset used for training, including its composition, biases, and limitations. This information helps stakeholders understand the data and its implications better. Model Cards, on the other hand, provide insights into the model itself, including its architecture, training process, and performance metrics. These documents facilitate a better understanding of the model's strengths, weaknesses, and potential ethical considerations.

By leveraging these methods, the transparency and accountability of the developed models can be enhanced. This is crucial in the context of AI-assisted embryo selection, as it promotes trust among clinicians, patients, and stakeholders involved in the decision-making process.

Bibliography

[1] T. Gomez, T. Fréour, and H. Mouchère, "Comparison of attention models and post-hoc explanation methods for embryo stage identification: a case study," May 2022, arXiv:2205.06546 [cs]. [Online]. Available: http://arxiv.org/abs/2205.06546 vii, 3, 8, 32, 33

[2] J. Boivin, L. Bunting, J. A. Collins, and K. G. Nygren, "International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care," *Human reproduction*, vol. 22, no. 6, pp. 1506–1512, 2007. 1

[3] A. C. Gore, V. A. Chappell, S. E. Fenton, J. A. Flaws, A. Nadal, G. S. Prins, J. Toppari, and R. T. Zoeller, "EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals," *Endocrine Reviews*, vol. 36, no. 6, pp. E1–E150, Dec. 2015. 1

[4] T. R. Segal and L. C. Giudice, "Before the beginning: environmental exposures and reproductive and obstetrical outcomes," *Fertility and Sterility*, vol. 112, no. 4, pp. 613–621, Oct. 2019. 1

[5] Y. Mio and K. Maeda, "Time-lapse cinematography of dynamic changes occurring during in vitro development of human embryos," *American journal of obstetrics and gynecology*, vol. 199, no. 6, pp. 660–e1, 2008. 1

[6] N. Nasiri and P. Eftekhari-Yazdi, "An overview of the available methods for morphological scoring of pre-implantation embryos in in vitro fertilization," *Cell Journal (Yakhteh)*, vol. 16, no. 4, p. 392, 2015. 1

[7] D. K. Gardner and D. Sakkas, "Assessment of embryo viability: the ability to select a single embryo for transfer–a review," *Placenta*, vol. 24 Suppl B, pp. S5–12, Oct. 2003. 1, 8

[8] Comisión Nacional de Reproducción Humana Asistida, 2021. [Online]. Available: https://cnrha.sanidad.gob.es/registros/actividades.htm 2, 43

[9] Z. Rosenwaks, "Artificial intelligence in reproductive medicine: a fleeting concept or the wave of the future?" *Fertility and Sterility*, vol. 114, no. 5, pp. 905–907, 2020. 2

[10] M. F. Kragh and H. Karstoft, "Embryo selection with artificial intelligence: how to evaluate and compare methods?" *Journal of Assisted Reproduction and Genetics*, vol. 38, no. 7, pp. 1675–1689, Jul. 2021. 2, 17, 43, 44, 45, 46

[11] Jefatura del Estado, "Ley 14/2006, de 26 de mayo, sobre técnicas de reproducción humana asistida," pp. 19 947–19 956, May 2006. [Online]. Available: https://www.boe.es/eli/es/l/2006/05/26/14 6

[12] A. Barbuscia, P. Martikainen, M. Myrskylä, H. Remes, E. Somigliana, R. Klemetti, and A. Goisis, "Maternal age and risk of low birth weight and premature birth in children conceived through medically assisted reproduction. evidence from finnish population registers," *Human Reproduction*, vol. 35, no. 1, pp. 212–220, 2020. 9, 12

[13] T. Gomez, M. Feyeux, J. Boulant, N. Normand, L. David, P. Paul-Gilloteaux, T. Fréour, and H. Mouchère, "A time-lapse embryo dataset for morphokinetic parameter prediction," *Data in Brief*, vol. 42, p. 108258, 2022. 9

[14] F. Kromp, R. Wagner, B. Balaban, V. Cottin, I. Cuevas-Saiz, C. Schachner, P. Fancsovits, M. Fawzy, L. Fischer, N. Findikli *et al.*, "An annotated human blastocyst dataset to benchmark deep learning architectures for in vitro fertilization," *Scientific Data*, vol. 10, no. 1, p. 271, 2023. 9

[15] [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj 10

[16] I. Dimitriadis, N. Zaninovic, A. C. Badiola, and C. L. Bormann, "Artificial intelligence in the embryology laboratory: a review," *Reproductive BioMedicine Online*, vol. 44, no. 3, pp. 435–448, Mar. 2022, publisher: Elsevier. [Online]. Available: https://www.rbmojournal.com/article/S1472-6483(21)00557-5/fulltext 12

[17] I. S. Isa, U. K. Yusof, and M. Mohd Zain, "Image Processing Approach for Grading IVF Blastocyst: A State-of-the-Art Review and Future Perspective of Deep Learning-Based Models," *Applied Sciences*, vol. 13, no. 2, p. 1195, Jan. 2023, number: 2 Publisher:

Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/13/2/1195 12

[18] E. M. Konstantinos Sfakianoudis, "Reporting on the Value of Artificial Intelligence in Predicting the Optimal Embryo for Transfer: A Systematic Review including Data Synthesis," *Biomedicines*, vol. 10, no. 3, p. 697, 2022. [Online]. Available: https://www.mdpi.com/2227-9059/10/3/697 12, 46

[19] C. M. Louis, A. Erwin, N. Handayani, A. A. Polim, A. Boediono, and I. Sini, "Review of computer vision application in in vitro fertilization: the application of deep learning-based computer vision technology in the world of IVF," *Journal of Assisted Reproduction and Genetics*, vol. 38, no. 7, pp. 1627–1639, Jul. 2021. 12

[20] K. Lundin and H. Park, "Time-lapse technology for embryo culture and selection," *Upsala Journal of Medical Sciences*, vol. 125, no. 2, pp. 77–84, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7720962/ 12

[21] N. Zaninovic and Z. Rosenwaks, "Artificial intelligence in human in vitro fertilization and embryology," *Fertility and Sterility*, vol. 114, no. 5, pp. 914–920, Nov. 2020, publisher: Elsevier. [Online]. Available: https://www.fertstert.org/article/S0015-0282(20)32399-2/fulltext 12

[22] E. I. Fernandez, A. S. Ferreira, M. H. M. Cecílio, D. S. Chéles, R. C. M. de Souza, M. F. G. Nogueira, and J. C. Rocha, "Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data," *Journal of Assisted Reproduction and Genetics*, vol. 37, no. 10, pp. 2359–2376, Oct. 2020. [Online]. Available: https://link.springer.com/10.1007/s10815-020-01881-9 12, 13

[23] J. Kim, J. Lee, and J. H. Jun, "Non-invasive evaluation of embryo quality for the selection of transferable embryos in human in vitro fertilization-embryo transfer," *Clinical and Experimental Reproductive Medicine*, vol. 49, no. 4, pp. 225–238, Dec. 2022. 13

[24] M. A. M. Afnan, C. Rudin, V. Conitzer, J. Savulescu, A. Mishra, Y. Liu, and M. Afnan, "Ethical Implementation of Artificial Intelligence to Select Embryos in In Vitro Fertilization," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA:

Association for Computing Machinery, Jul. 2021, pp. 316–326. [Online]. Available: https://doi.org/10.1145/3461702.3462589 14, 49, 52, 53

[25] S. Tamir, "Artificial intelligence in human reproduction: charting the ethical debate over AI in IVF," *AI and Ethics*, Sep. 2022. [Online]. Available: https://doi.org/10.1007/s43681-022-00216-x 14, 49, 50, 51, 52, 53

[26] V. Rolfes, U. Bittner, H. Gerhards, J.-S. Krüssel, T. Fehm, R. Ranisch, and H. Fangerau, "Artificial Intelligence in Reproductive Medicine An Ethical Perspective," *Geburtshilfe und Frauenheilkunde*, vol. 83, no. 1, pp. 106–115, Jan. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9833891/ 14, 49, 52, 53

[27] T. Chen, Y. Cheng, J. Wang, Z. Yang, W. Zheng, D. Z. Chen, and J. Wu, "Automating blastocyst formation and quality prediction in time-lapse imaging with adaptive key frame selection," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*. Springer, 2022, pp. 445–455. 18, 25, 26, 41, 46

[28] X. Xie, P. Yan, F.-Y. Cheng, F. Gao, Q. Mai, and G. Li, "Early prediction of blastocyst development via time-lapse video analysis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5. 18, 25, 26, 42

[29] Q. Liao, Q. Zhang, X. Feng, H. Huang, H. Xu, B. Tian, J. Liu, Q. Yu, N. Guo, Q. Liu *et al.*, "Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring," *Communications biology*, vol. 4, no. 1, p. 415, 2021. 18, 19, 25, 26, 28, 29, 38, 40

[30] Y. Kan-Tor, N. Zabari, I. Erlich, A. Szeskin, T. Amitai, D. Richter, Y. Or, Z. Shoham, A. Hurwitz, I. Har-Vardi *et al.*, "Automated evaluation of human embryo blastulation and implantation potential using deep-learning," *Advanced Intelligent Systems*, vol. 2, no. 10, p. 2000080, 2020. 18, 20, 25, 26, 28, 36, 38, 40, 42, 48

[31] J. T. Lassen, M. F. Kragh, J. Rimestad, M. N. Johansen, and J. Berntsen, "Development and validation of deep learning based embryo selection across multiple days of transfer," 2022. 19, 27, 30

[32] S. M. Diakiw, J. M. M. Hall, M. VerMilyea, A. Y. X. Lim, W. Quangkananurug, S. Chanchamroen, B. Bankowski, R. Stones,

A. Storr, A. Miller, G. Adaniya, R. van Tol, R. Hanson, J. Aizpurua, L. Giardini, A. Johnston, T. Van Nguyen, M. A. Dakka, D. Perugini, and M. Perugini, "An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos," *Reproductive BioMedicine Online*, vol. 45, no. 6, pp. 1105–1117, Dec. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1472648322005235 19, 28, 29, 33, 34, 45

[33] J. Berntsen, J. Rimestad, J. T. Lassen, D. Tran, and M. F. Kragh, "Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences," *PloS One*, vol. 17, no. 2, p. e0262661, 2022. 19, 27, 28, 36, 41, 42, 47

[34] N. Enatsu, I. Miyatsuka, L. M. An, M. Inubushi, K. Enatsu, J. Otsuki, T. Iwasaki, S. Kokeguchi, and M. Shiotani, "A novel system based on artificial intelligence for predicting blastocyst viability and visualizing the explanation," *Reproductive Medicine and Biology*, vol. 21, no. 1, p. e12443, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rmb2.12443. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rmb2.12443 19, 21, 28, 29, 31, 33, 34, 37, 39, 46, 51

[35] A. Chavez-Badiola, A. Flores-Saiffe-Farías, G. Mendizabal-Ruiz, A. J. Drakeley, and J. Cohen, "Embryo ranking intelligent classification algorithm (erica): artificial intelligence clinical assistant predicting embryo ploidy and implantation," *Reproductive BioMedicine Online*, vol. 41, no. 4, pp. 585–593, 2020. 19, 27, 28, 29, 38, 45

[36] C. L. Bormann, M. K. Kanakasabapathy, P. Thirumalaraju, R. Gupta, R. Pooniwala, H. Kandula, E. Hariton, I. Souter, I. Dimitriadis, L. B. Ramirez *et al.*, "Performance of a deep learning based neural network in the selection of human blastocysts for implantation," *Elife*, vol. 9, p. e55301, 2020. 20, 28, 29, 36, 38, 49

[37] D. H. Silver, M. Feder, Y. Gold-Zamir, A. L. Polsky, S. Rosentraub, E. Shachor, A. Weinberger, P. Mazur, V. D. Zukin, and A. M. Bronstein, "Data-driven prediction of embryo implantation probability using ivf time-lapse imaging," *arXiv preprint arXiv:2006.01035*, 2020. 20, 28, 29, 36

[38] M. VerMilyea, J. M. M. Hall, S. M. Diakiw, A. Johnston, T. Nguyen, D. Perugini, A. Miller, A. Picou, A. P. Murphy, and M. Perugini, "Development of an artificial intelligence-based assessment model for prediction of embryo

viability using static images captured by optical light microscopy during IVF," *Human Reproduction (Oxford, England)*, vol. 35, no. 4, pp. 770–784, Apr. 2020. 20, 28, 29, 36, 47

[39] D. Tran, S. Cooke, P. J. Illingworth, and D. K. Gardner, "Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer," *Human reproduction*, vol. 34, no. 6, pp. 1011–1018, 2019. 20, 28, 29, 36, 43, 47

[40] Q. Cao, S. S. Liao, X. Meng, H. Ye, Z. Yan, and P. Wang, "Identification of viable embryos using deep learning for medical image," in *Proceedings of the 2018 5th international conference on bioinformatics research and applications*, 2018, pp. 69–72. 20, 28, 29

[41] B. Huang, S. Zheng, B. Ma, Y. Yang, S. Zhang, and L. Jin, "Using deep learning to predict the outcome of live birth from more than 10,000 embryo data," *BMC pregnancy and childbirth*, vol. 22, no. 1, p. 36, Jan. 2022. 21, 30, 31, 43, 46

[42] Y. Sawada, T. Sato, M. Nagaya, C. Saito, H. Yoshihara, C. Banno, Y. Matsumoto, Y. Matsuda, K. Yoshikai, T. Sawada, N. Ukita, and M. Sugiura-Ogasawara, "Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth," *Reproductive Biomedicine Online*, vol. 43, no. 5, pp. 843–852, Nov. 2021. 21, 30, 31, 33, 34, 40, 43, 49, 51

[43] Y. Miyagi, T. Habara, R. Hirata, and N. Hayashi, "Predicting a live birth by artificial intelligence incorporating both the blastocyst image and conventional embryo evaluation parameters," *Artif Intell Med Imaging*, vol. 1, no. 3, pp. 94–107, 2020. 21, 30, 31, 37

[44] J. Silva-Rodriguez, A. Colomer, M. Meseguer, and V. Naranjo, "Predicting the Success of Blastocyst Implantation from Morphokinetic Parameters Estimated through CNNs and Sum of Absolute Differences," in *2019 27th European Signal Processing Conference (EUSIPCO)*. A Coruna, Spain: IEEE, Sep. 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8902520/ 21, 30, 31, 38, 41, 46

[45] Y. Miyagi, T. Habara, R. Hirata, and N. Hayashi, "Feasibility of deep learning for predicting live birth from a blastocyst image in patients classified by age," *Reproductive Medicine and Biology*, vol. 18, no. 2, pp. 190–203, Apr. 2019. 21, 30, 31, 36

[46] J. Theilgaard Lassen, M. Fly Kragh, J. Rimestad, M. Nygård Johansen, and J. Berntsen, "Development and validation of deep learning based embryo

selection across multiple days of transfer," *Scientific Reports*, vol. 13, no. 1, p. 4235, 2023. 28, 36, 41, 42, 47, 48

[47] L. Alegre, L. Bori, M. de los Ángeles Valera, M. F. G. Nogueira, A. S. Ferreira, J. C. Rocha, and M. Meseguer, "First application of artificial neuronal networks for human live birth prediction on geri time-lapse monitoring system blastocyst images," *Fertility and Sterility*, vol. 114, no. 3, p. e140, 2020. 31

[48] M. Meseguer, C. Hickman, L. B. Arnal, L. Alegre, M. Toschi, R. Del Gallego, and J. C. Rocha, "Is there any room to improve embryo selection? artificial intelligence technology applied for ive birth prediction on blastocysts," *Fertility and Sterility*, vol. 112, no. 3, p. e77, 2019. 31

[49] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 1135–1144, 2016. 32

[50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Red Hook, NY, USA, p. 47684777, 2017. 32

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," pp. 618–626, 2017. 32

[52] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," Berlin, Heidelberg, pp. 404–417, 2006. 33

[53] A. Kallipolitis, M. Tziomaka, D. Papadopoulos, and I. Maglogiannis, "Explainable computer vision analysis for embryo selection on blastocyst images," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Sep. 2022, pp. 1–4, iSSN: 2641-3604. 33, 34, 51

[54] C. Wu, L. Fu, Z. Tian, J. Liu, J. Song, W. Guo, Y. Zhao, D. Zheng, Y. Jin, D. Yi, and X. Jiang, "LWMA-Net: Light-weighted morphology attention learning for human embryo grading," *Computers in Biology and Medicine*, vol. 151, p. 106242, Dec. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482522009507 33, 49, 51

[55] M. Arsalan, A. Haider, S. W. Cho, Y. H. Kim, and K. R. Park, "Human Blastocyst Components Detection Using Multiscale Aggregation

Semantic Segmentation Network for Embryonic Analysis," *Biomedicines*, vol. 10, no. 7, p. 1717, Jul. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9313331/ 33, 34, 51

[56] A. Sharma, M. H. Stensen, E. Delbarre, T. B. Haugen, and H. L. Hammer, "Explainable Artificial Intelligence for Human Embryo Cell Cleavage Stages Analysis," in *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*. Newark NJ USA: ACM, Jun. 2022, pp. 1–8. [Online]. Available: https://dl.acm.org/doi/10.1145/3512731.3534206 33, 34, 51

[57] E. Payá, L. Bori, A. Colomer, M. Meseguer, and V. Naranjo, "Automatic characterization of human embryos at day 4 post-insemination from time-lapse imaging using supervised contrastive learning and inductive transfer learning techniques," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106895, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722002772 33, 34

[58] S. Wang, C. Zhou, D. Zhang, L. Chen, and H. Sun, "A Deep Learning Framework Design for Automatic Blastocyst Evaluation With Multifocal Images," *IEEE Access*, vol. 9, pp. 18 927–18 934, 2021, conference Name: IEEE Access. 33, 38, 42, 51

[59] P. Thirumalaraju, M. K. Kanakasabapathy, C. L. Bormann, R. Gupta, R. Pooniwala, H. Kandula, I. Souter, I. Dimitriadis, and H. Shafiee, "Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality," May 2020, arXiv:2005.10912 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2005.10912 33, 34, 51

[60] A. Chavez-Badiola, A. Flores-Saiffe Farias, G. Mendizabal-Ruiz, R. Garcia-Sanchez, A. J. Drakeley, and J. P. Garcia-Sandoval, "Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning," *Scientific reports*, vol. 10, no. 1, p. 4394, 2020. 36

[61] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. 38

[62] T.-J. Chen, W.-L. Zheng, C.-H. Liu, I. Huang, H.-H. Lai, and M. Liu, "Using deep learning with large dataset of microscope images to develop

an automated embryo grading system," *Fertility & Reproduction*, vol. 1, no. 01, pp. 51–56, 2019. 44

[63] K. G. Moons, D. G. Altman, J. B. Reitsma, J. P. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration," *Annals of internal medicine*, vol. 162, no. 1, pp. W1–W73, 2015. 45, 46, 47

[64] P. Ala-Pietilä et al., "Ethics Guidelines for Trustworthy AI," Dec. 2018. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation 50, 56

[65] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi, "Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–9. 51, 53

[66] A. Levy, M. Agrawal, A. Satyanarayan, and D. Sontag, "Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13. 51

[67] Y. Harada, S. Katsukura, R. Kawamura, and T. Shimizu, "Effects of a differential diagnosis list of artificial intelligence on differential diagnoses by physicians: an exploratory analysis of data from a randomized controlled study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 5562, 2021. 51

[68] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004. 51

[69] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, "" hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making," *Proceedings of the ACM on Human-computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019. 51

[70] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, J. Adebayo, M. D. Li, and J. Kalpathy-Cramer, "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021. 51

[71] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021. 51

[72] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021. 56, 62

[73] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229. 56, 62

[74] P. Khosravi, E. Kazemi, Q. Zhan, J. E. Malmsten, M. Toschi, P. Zisimopoulos, A. Sigaras, S. Lavery, L. A. D. Cooper, C. Hickman, M. Meseguer, Z. Rosenwaks, O. Elemento, N. Zaninovic, and I. Hajirasouliha, "Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization," *npj Digital Medicine*, vol. 2, no. 1, p. 21, Apr. 2019. [Online]. Available: https://www.nature.com/articles/s41746-019-0096-y

[75] K. G. Moons, R. F. Wolff, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "Probast: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration," *Annals of internal medicine*, vol. 170, no. 1, pp. W1–W33, 2019.

[76] B. Huang, W. Tan, Z. Li, and L. Jin, "An artificial intelligence model (euploid prediction algorithm) can predict embryo ploidy status based on time-lapse data," *Reproductive biology and endocrinology: RB&E*, vol. 19, no. 1, p. 185, Dec. 2021.

[77] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*. Hong Kong, China: ACM Press, 2011, p. 177. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1935826.1935863

[78] R. Becker, D. Chokoshvili, G. Comandé, E. S. Dove, A. Hall, C. Mitchell, F. Molnár-Gábor, P. Nicolàs, S. Tervo, and A. Thorogood, "Secondary Use of Personal Health Data: When Is It Further Processing Under the GDPR, and What Are the Implications for Data Controllers?" *European Journal of Health Law*, vol. -1,

no. aop, pp. 1–29, Aug. 2022, publisher: Brill Nijhoff. [Online]. Available: https://brill.com/view/journals/ejhl/aop/article-10.1163-15718093-bja10094/article-10.1163-15718093-bja10094.xml

[79] G. Comandè and G. Schneider, "Differential Data Protection Regimes in Data-Driven Research: Why the GDPR is More Research-Friendly Than You Think," *German Law Journal*, vol. 23, no. 4, pp. 559–596, May 2022, publisher: Cambridge University Press.

[80] "A time-lapse embryo dataset for morphokinetic parameter prediction | Elsevier Enhanced Reader."

[81] B. M. Rasmussen, "KIDScore Decision Support Tool."

[82] M. de Araujo, "The Ethics of Genetic Cognitive Enhancement: Gene Editing or Embryo Selection?" *Philosophies*, vol. 5, no. 3, p. 20, Sep. 2020, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2409-9287/5/3/20

[83] "O-098 Embryo selection using Artificial Intelligence (AI): Epistemic and ethical considerations | Human Reproduction | Oxford Academic." [Online]. Available: https://academic.oup.com/humrep/article/36/Supplement_1/deab125.034/6343660?login=false

[84] L. Bori, R. Maor, F. Meseguer, I. Kottel, D. S. Seidman, D. Gilboa, and M. Meseguer, "ARTIFICIAL INTELLIGENCE IS MOVING CLOSER TO REPRODUCTIVE MEDICINE: PREDICTION OF BLASTULATION AND EMBRYO IMPLANTATION," *Fertility and Sterility*, vol. 116, no. 3, p. e154, Sep. 2021, publisher: Elsevier. [Online]. Available: https://www.fertstert.org/article/S0015-0282(21)01017-7/fulltext

[85] J. Barnes, M. Brendel, V. R. Gao, S. Rajendran, J. Kim, Q. Li, J. E. Malmsten, J. T. Sierra, P. Zisimopoulos, A. Sigaras, P. Khosravi, M. Meseguer, Q. Zhan, Z. Rosenwaks, O. Elemento, N. Zaninovic, and I. Hajirasouliha, "A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study," *The Lancet Digital Health*, vol. 5, no. 1, pp. e28–e40, Jan. 2023, publisher: Elsevier. [Online]. Available: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00213-8/fulltext

[86] S. N. Patil, D. U. Wali, and D. M. K. Swamy, "Deep Learning Techniques for Automatic Classification and Analysis of Human in Vitro Fertilized (IVF) embryos." 2018.

[87] M. F. Kragh, J. Rimestad, J. Berntsen, and H. Karstoft, “Automatic grading of human blastocysts from time-lapse imaging,” *Computers in Biology and Medicine*, vol. 115, p. 103494, Dec. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519303609

[88] D. Cimadomo, V. Chiappetta, F. Innocenti, G. Saturno, M. Taggi, A. Marconetto, V. Casciani, L. Albricci, R. Maggiulli, G. Coticchio, A. Ahlström, J. Berntsen, M. Larman, A. Borini, A. Vaiarelli, F. M. Ubaldi, and L. Rienzi, “Towards Automation in IVF: Pre-Clinical Validation of a Deep Learning-Based Embryo Grading System during PGT-A Cycles,” *Journal of Clinical Medicine*, vol. 12, no. 5, p. 1806, Feb. 2023.

[89] N. Rostamzadeh, D. Mincu, S. Roy, A. Smart, L. Wilcox, M. Pushkarna, J. Schrouff, R. Amironesei, N. Moorosi, and K. Heller, “Healthsheet: development of a transparency artifact for health datasets,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1943–1961.

[90] G. D. Adamson and R. J. Norman, “Why are multiple pregnancy rates and single embryo transfer rates so different globally, and what do we do about it?” *Fertility and Sterility*, vol. 114, no. 4, pp. 680–689, Oct. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0015028220322032

[91] Q. Zhao, E. Adeli, and K. M. Pohl, “Training confounder-free deep learning models for medical applications,” *Nature Communications*, vol. 11, no. 1, p. 6010, Nov. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-020-19784-9

[92] A. D’Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman *et al.*, “Underspecification presents challenges for credibility in modern machine learning,” *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 10 237–10 297, 2022.

[93] X. Xie, P. Yan, F.-Y. Cheng, F. Gao, Q. Mai, and G. Li, “Early Prediction of Blastocyst Development via Time-Lapse Video Analysis,” 2022, iSSN: 1945-7928.

[94] Y. Tokuoka, T. Yamada, D. Mashiko, Z. Ikeda, T. Kobayashi, K. Yamagata, and A. Funahashi, “An explainable deep learning-based algorithm with

an attention mechanism for predicting the live birth potential of mouse embryos," *Artificial Intelligence in Medicine*, vol. 134, 2022.

[95] L. Bori, R. Maor, F. Meseguer, I. Kottel, D. S. Seidman, D. Gilboa, and M. Meseguer, "Artificial intelligence is moving closer to reproductive medicine: Prediction of blastulation and embryo implantation," *Fertility and Sterility*, vol. 116, no. 3, p. e154, 2021.

[96] R. Valencia, A. F. S. Farías, G. Mendizabal, A. Chavez-Badiola, and F. J. A. Gomez, "Towards an explainable artificial intelligence to predict blastocyst formation potential from single oocyte images," *Fertility and Sterility*, vol. 118, no. 4, pp. e337–e338, 2022.

[97] G. Marvin and M. G. R. Alarm, "An Explainable Lattice based Fertility Treatment Outcome Prediction Model for TeleFertility," in *2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*. Dhaka, Bangladesh: IEEE, Dec. 2021, pp. 64–68. [Online]. Available: https://ieeexplore.ieee.org/document/9893623/

[98] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," 2018.

[99] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[100] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention Branch Network: Learning of Attention Mechanism for Visual Explanation," Long Beach, CA, USA, pp. 10 697–10 706, 2019.

[101] G. D. Adamson, J. d. Mouzon, G. M. Chambers, F. Zegers-Hochschild, R. Mansour, O. Ishihara, M. Banker, and S. Dyer, "International Committee for Monitoring Assisted Reproductive Technology: world report on assisted reproductive technology, 2011," *Fertility and Sterility*, vol. 110, no. 6, pp. 1067–1080, 2018.