Computers in Biology and Medicine 150 (2022) 106160

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Modeling three sources of uncertainty in assisted reproductive technologies with probabilistic graphical models

Jerónimo Hernández-González [a,*], Olga Valls [a], Adrián Torres-Martín [b], Jesús Cerquides [c]

[a] Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), 08007 Barcelona, Spain
[b] Department of Information and Communications Engineering, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain
[c] Artificial Intelligence Research Institute (IIIA-CSIC), 08193 Bellaterra, Spain

ARTICLE INFO

Keywords:
Assisted reproductive technologies
Embryo selection
Machine learning
Probabilistic graphical models
Expectation–Maximization

ABSTRACT

Embryo selection is a critical step in assisted reproduction: good selection criteria are expected to increase the probability of inducing a pregnancy. Machine learning techniques have been applied for implantation prediction or embryo quality assessment, which embryologists can use to make a decision about embryo selection. However, this is a highly uncertain real-world problem, and current proposals do not model always all the sources of uncertainty.

We present a novel probabilistic graphical model that accounts for three different sources of uncertainty, the standard embryo and cycle viability, and a third one that represents any unknown factor that can drive a treatment to a failure in otherwise perfect conditions. We derive a parametric learning method based on the Expectation–Maximization strategy, which accounts for uncertainty issues.

We empirically analyze the model within a real database consisting of 604 cycles (3125 embryos) carried out at Hospital Donostia (Spain). Embryologists followed the protocol of the Spanish Association for Reproduction Biology Studies (ASEBIR), based on morphological features, for embryo selection. Our model predictions are correlated with the ASEBIR protocol, which validates our model. The benefits of accounting for the different sources of uncertainty and the importance of the cycle characteristics are shown. Considering only transferred embryos, our model does not further discriminate them as implanted or failed, suggesting that the ASEBIR protocol could be understood as a thorough summary of the available morphological features.

1. Introduction

Assisted reproductive technologies (ARTs) are a set of invasive medical techniques that attempt to induce a pregnancy. Each trial of treatment is known as a cycle. The woman first follows a treatment of ovarian stimulation for several weeks to induce the development of multiple follicles with a large number of oocytes. Then, oocytes are retrieved and fertilized, and the resulting embryos are cultured for several days. Finally, clinicians need to select which embryos are transferred to the woman's uterus [1]. This process is physically and psychologically tough, especially for women, and success is not guaranteed. The Spanish Society of Fertilization (SEF) reported in 2018 that only 35.6% of the ART cycles succeeded (ended up in pregnancy) [2]. The probability of success can be improved by increasing the number of transferred embryos [3], but this also leads to higher multiple-birth rates, which is considered risky for both mother and fetuses [3,4]. Thus, many countries restrict the number of embryos that can be transferred (e.g., Spanish law limits it to 3). Therefore, the selection of the most viable embryos is a critical step to optimize the probability of pregnancy.

Embryo selection is a complex and partially subjective task. The evaluation of embryos is based mainly on the evolution of their morphological characteristics. The protocol of the Spanish Association for Reproduction Biology Studies (ASEBIR) [1], the criteria of reference in Spain, classifies embryos into an ordinal scale (from A –high-quality embryos– to D –low-quality embryos–) using morphological criteria and posed a unified protocol to address the lack of consensus in embryo quality assessment [5].

In recent years, machine learning (ML) techniques have been used to assist embryologists in embryo selection and pregnancy prediction [6–9]. Most of them rely on supervised classification and require complete and fully labeled training data. That is, we would need to know, for each embryo in our training dataset, its viability to induce a pregnancy. However, in ARTs, viability can only be determined after transference by the occurrence of embryo implantation. Moreover, for a transfer of

* Corresponding author.
E-mail address: jeronimo.hernandez@ub.edu (J. Hernández-González).

https://doi.org/10.1016/j.compbiomed.2022.106160
Received 29 April 2022; Received in revised form 8 September 2022; Accepted 1 October 2022
Available online 5 October 2022

multiple embryos, current techniques are unable to identify individually which embryo(s) implanted. This implies that many embryos are not (fully) labeled in ART data samples, and previous works usually discarded all the embryos lacking a full labeling. Nowadays, specific methods [8] have been proposed to learn also using information from cycles with partial implantation (not all the transferred embryos were implanted).

All current methods, ML-based or not, use a combination of embryos and cycles descriptive characteristics to predict embryo implantation. Yet, there exists a recurrent situation in assisted reproduction units: apparently viable cycles, using embryos allegedly viable, do not succeed. The repeated occurrence of this type of failure suggests that there exist still unknown factors which also determine cycle success.

In this paper, we propose a novel probabilistic graphical model that works under the assumption of independence between embryo and cycle viability, and accounts for a third source of uncertainty corresponding to unknown factors that can lead a cycle to fail. We have derived a learning algorithm specifically for this model based on the Expectation–Maximization (EM) strategy, given the context of partially labeled data and latent variables. We use two probabilistic classifiers to approximate the probabilistic distribution for embryo and cycle viability given their respective descriptive features.

We perform a thorough experimental validation of the model using real data. It is compared with several baseline approaches designed in an incremental way in order to test different working hypotheses. We also test the importance for embryo implantation prediction of the cycles features, as well as the relationship of the predictions of our model with the ASEBIR protocol. A preliminary version of this last part of the empirical validation was presented in [10]. The results show the ability of our model to learn and take advantage of all the available information. Its behavior is in line with the ASEBIR score, which validates our model.

The rest of the paper is organized as follows. First of all, the state of the art is reviewed. Then, we describe the real data available for this study and the proposed model, as well as the learning technique derived for it. In Section 4, we discuss a complete empirical evaluation of our model against several baseline techniques. The paper finishes drawing conclusions and future work.

2. State of the art

ART treatments are complex processes involving maternal hormonal changes, immune responses, and maturational events in the embryo. A treatment can fail when these events are not synchronized [11]. Despite the great improvements in ovarian stimulation protocols and fertilization procedures, implantation rates per embryo remain at approximately 15% and many patients experience multiple failed attempts [12]. Recurrent implantation failure (RIF) is a condition resulting from repetitive unsuccessful ART cycles [13], and it provides evidence of the existence of still unknown factors that affect ART success.

All this has provided an ideal context for the application of ML methods. Since more than 20 years ago, ML techniques provide the standardized and efficient tools demanded in laboratories for evaluating the different processes in ARTs: from embryo selection to assessing patient reproductive potential, or individualizing stimulation protocols [14]. Since the popularization of infertility treatments, many works have focused on the problem of ART outcome prediction and one of its critical steps: the selection of embryos [15–17]. This subfield has rapidly evolved in the last decade due to technological advances. In the classical scenario, embryologists collect the most relevant morphological traits of embryos by visual inspection of them via microscopes [18]. More recently, the use of (static) photographs of the embryos enabled the use of automated image processing techniques [19,20]. Current embryo incubators incorporate cameras that allow embryologists to inspect the whole evolution of the embryos through time-lapse videos. This data is being fed directly to ML models for embryo viability prediction [21]. Indeed, these sources of data are complementary and can be combined in a single ML method [22,23].

The technological breakthroughs have brought novel machine learning methods too, which have been applied to ARTs. Standard artificial intelligence techniques have been used, such as ranking algorithms [24], statistical models, ensemble techniques, neural networks [25], Bayesian networks [6,8,26], Support Vector Machines [27], classification and regression trees, logistic regression, case-based reasoning systems, etc. [15,17]. More recently, deep learning methods [9,28] have been used to analyze the vast amount of data coming from time-lapse incubators. ML techniques are of great interest since traditional morphokinetic grading systems can be subjective and variable. It is generally agreed that ML methods are promising for the ART community but still require further validation [29]. For example, fully automated (time-lapse imaging) approaches require costly equipment and have not demonstrated sufficient predictive ability yet [30]. Moreover, there exist doubts about the deployment of these systems in the medical domain, regarding technological and ethical aspects. Recently, Müller et al. [31] proposed a list of ten principles for designing ML-based decision support systems: an ethical system should be transparent, explainable, fair, repeatable, under responsibility and monitored by a physical person, and its suggestions must imply no human harm.

Most of these ML works take the standard supervised classification approach, which requires completely labeled datasets. However, labeling all the embryos is not always possible: in a cycle where not all the transferred embryos get implanted, the use of current medical techniques allows for knowing how many embryos got implanted, but not to know exactly which embryo did. Many previous works [9,32,33] directly disregard the embryos from these partially observed cycles. Morales et al. [26] proposed, to circumvent this issue, joining the descriptive vectors of all embryos in each cycle and learning to predict a pregnancy. Hernández-González et al. [8] reformulated the task as a weakly supervised learning problem, and learned using all the embryos and the available information of supervision (label proportions or counts of implanted embryos per cycle).

Another widespread approach is the embryo–uterine model (EU), introduced by Speirs et al. [34] and later extended by Zhou and Weinberg [35]. It assumes that, for a pregnancy to happen, both a fertile patient (receptive uterus) and a viable embryo are required. Two separate modules (embryo [E] and uterus [U]) compose it: the probability of implantation is predicted as the product of the probabilities given by both submodules. These models suffer from even harder issues of partial observability: if a cycle fails and no embryo implants, one cannot know if the embryos were not viable, if the cycle was not fertile, or both. Roberts [18] addressed this via the Expectation–Maximization (EM) algorithm, and Corani et al. [6] used a Bayesian network trained with an averaging approach as an alternative to MAP estimation using a very limited set of descriptive features for cycles and embryos. Roberts and Stylianou [36] used an EU model to try to assess other unknown factors that might be related to a given patient when they undergo several ART cycles.

3. Materials and methods

3.1. Data

The database, originally presented by Hernández-González et al. [8], was collected by the Unit of Assisted Reproduction of the Hospital Donostia (Spain) from January 2013 to June 2015. In total, 604 cycles were carried out, compiling a total number of 3125 embryos. Each cycle has a certain number of embryos associated, only some of which were actually transferred. As detailed in Table 1, in this period 412 cycles failed to induce a pregnancy (839 embryos), and only in 57 cycles did all the transferred embryos (108) result implanted. In the remaining

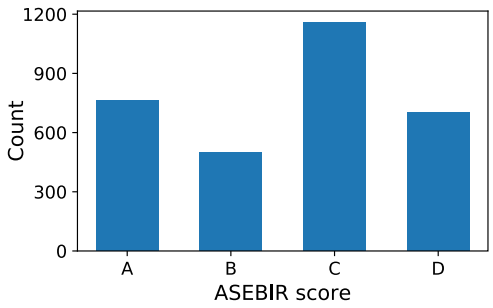


Fig. 1. Embryo counts for each of the categories of the ASEBIR scoring system [1].

135 cycles, only a subset of the 307 transferred embryos were implanted. This last subset is of relevance in our analysis, as we cannot determine the real fate of each embryo individually (it is not possible to know which specific embryos are the ones implanted). Among all the cycles, up to 1871 embryos were not selected for transfer. The criteria for limiting the number of embryos to transfer goes from the low quality of the embryos to legal restrictions (in Spain, the maximum number of embryos that can be transferred in a single trial is 3).

Each cycle is described by 25 features including characteristics of the patients (female and male) and stimulation procedure. Moreover, summary variables of the associated embryos are provided (e.g., cycle's fertility rate, i.e., the proportion of oocytes successfully fertilized). Each embryo is described by 20 features, mainly morphological characteristics at different stages of development (up to 48 h after fertilization, when transference was carried out). Appendix A details the descriptive features of both subsets. In practice, only informative variables were considered. We used one-hot encoding to transform categorical into numeric features. All of them were then standardized (centered and scaled to unit variance). After all, 36 features for cycles and 25 for embryos were left.

A key feature is *success rate* ($\in [0,1]$), which indicates the percentage of transferred embryos in the cycle that implanted. Note that this is the ultimate information we would like our models to predict. The value 0 indicates that all the embryos of the cycle failed to implant, 1 that all of them implanted, and any value in the interval $(0,1)$ indicates the proportion of implanted (and failed) embryos. This latter case is directly related to the aforementioned problem of partially observed data: we cannot know the identity of the implanted embryos (we do not know their actual outcome) in the cycle, we only know that some of them were implanted.

For each embryo, we also have a quality score (A–D, from high to low-quality embryos) given by embryologists according to the ASEBIR protocol [1], which assigns each embryo to a category based on its morphological characteristics. The distribution of embryos among categories is rather balanced, although category C (mid-low quality) stands out (see Fig. 1). This quality score is a decisive factor in the selection process performed by embryologists, as can be seen in Fig. 2(a).

Ideally, there should be a clear difference in the implantation rate of embryos graded in different categories. We display in Fig. 2(b) the fraction of transferred embryos with different outcomes for each quality category. It can be observed that there is a small signal: as the quality score increases, the proportion of non-implanted embryos decreases. Being aware of the difficulty of this problem, these numbers support the effectiveness of the ASEBIR protocol to indicate implantation.

3.2. A probabilistic implantation model for ART

In this paper, we propose a probabilistic model that comprehensively accounts for three different sources of uncertainty in the ART problem, which is presented in this section. Later on, we present its learning method that uses all the available information, even the partial label information from cycles where not all the embryos were implanted.

We model the problem of ART by means of a probabilistic graphical model (PGM) [37], which grounds on a solid mathematical background. A directed acyclic graph is used to encode a set of conditional independencies between the random variables, and the join distribution factorizes as the product of conditional probability distributions for each variable given its parents. Given a fixed structure, the model parameters can be estimated from data.

The proposed model takes into account three sources of uncertainty related to the success of an ART procedure, namely the viability of embryos and cycles, and other unknown factors.

The viability of the embryo. A widely accepted assumption in ART is that the individual characteristics of an embryo ($\boldsymbol{x}_e$) are relevant in order to predict the probability of such embryo implanting in the uterus. According to the provided data, an embryo's viability is assumed to be related to its morphological traits. Here, the distribution

$$p(w_e \mid \boldsymbol{x}_e; \alpha) \tag{1}$$

measures the probability of the embryo to implant in a "perfect cycle" (fully fertile patient), where $\boldsymbol{x}_e$ represents the descriptive characteristics of embryos. We will model this distribution with a probabilistic classifier.

The viability of the cycle. Another common assumption is that the individual patient features and the undergone stimulation treatment exert an influence on the likelihood of her fertility potential. This is how we define cycle viability. The distribution

$$p(r_c \mid \boldsymbol{v}_c; \beta) \tag{2}$$

assesses how the descriptive characteristics of the cycle, $\boldsymbol{v}_c$, influence fertility potential. We will model this distribution through a probabilistic classifier too.

These two components form the classical embryo–uterine modeling approach. It implies that we assume that the fertility potential of the patient is statistically independent from the embryo characteristics. This is a practical assumption, but highly unlikely when the patient's own oocytes are used, as was the case in this study.

Other unknown factors. There is consensus in the ART scientific literature that there are still unknown factors that (partially) determine the outcome of an ART treatment, like those provoking recurrent implantation failure [38]. We model this uncertainty by means of a Bernoulli distribution, with parameter $\theta_1 \in [0,1]$. The implantation of a viable embryo in a fertile cycle follows a distribution

$$i_e^c \sim Bernoulli(\theta_{r_c \cdot w_e \cdot s_e^c}) \tag{3}$$

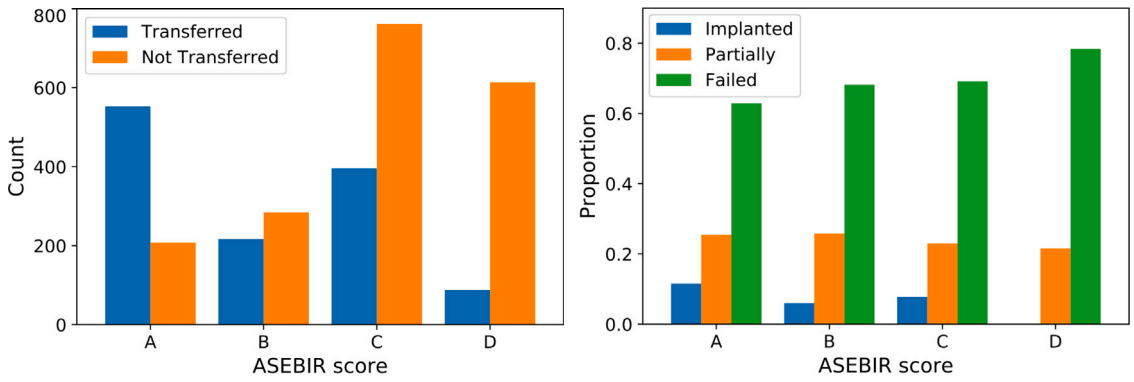
where s_e^c is 1 if embryo e was transferred in cycle c, and 0 otherwise. θ_1 is the probability that in a cycle that has been properly configured ($r_c = 1$), a viable embryo ($w_e = 1$), selected for transfer ($s_e^c = 1$), gets implanted. Ideally, there would be no such unknown factor and $\theta_1 = 1$. For modeling convenience, we use a second Bernoulli with fixed parameter $\theta_0 = 0$, which tells that there will not be implantation if any r.v. w_e, r_c or s_e^c is 0 (no viable embryo or cycle, or embryo not transferred).

Finally, the number of embryos implanted in cycle c, i.e., the outcome y_c, is deterministically assessed as:

$$y_c = \sum_{e \in E^c} i_e^c. \tag{4}$$

Note that depending on the practice of the specific ART unit, more than one transference could be carried out for the same cycle. Following the practice of our Unit of reference (and as reflected in the data), here we only consider the case where a single transfer of one or more embryos is carried out in each cycle.

The graphical structure of our model is shown in Fig. 3. The observed variables are shadowed ($\boldsymbol{x}_e, \boldsymbol{v}_c, s_e^c, y_c$), whereas white nodes



(a) Number of embryos, according to use (transferred or not) and ASEBIR category.

(b) Proportion of transferred embryos with each fate, for all ASEBIR categories.

Fig. 2. Distributions of embryos in the database regarding the ASEBIR scoring system [1].

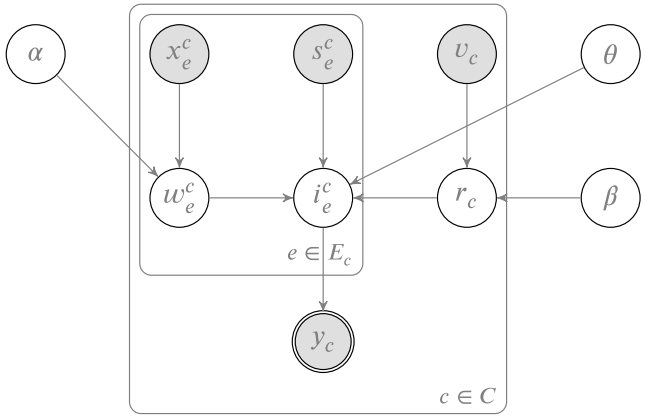


Fig. 3. Graphical description of the model. Shadowed nodes represent observed variables. Double lines denote deterministic variables.

Table 1
Number of cycles and embryos, separated by use (transferred or not) and success (pregnancy or not, embryo implanted or not).

Cycles		Embryos	
Unsuccessful	412	839	All failed to implant
Successful	135	307	Some failed to implant
Successful	57	108	All implanted
Total no.	604	1254	Transferred (*Subtotal no.*)
		1871	Not transferred
		3125	*Total no.*

(w_e, r_c, i_e^c) represent latent variables, the value of which need to be inferred. In certain cases, the values of some of these latter r.v. can be known. Finally, α, β, θ are the hyper-parameters of the cycles' and embryo's classifiers (Eqs. (1) and (2)) and of the Bernoulli distribution (Eq. (3)). All the notation is summarized in Table 2.

The joint probability distribution of the model is

$$p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{v}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{i}, \boldsymbol{y}; \alpha, \beta, \theta) = p(\boldsymbol{w}|\boldsymbol{x}; \alpha)p(\boldsymbol{x})p(\boldsymbol{r}|\boldsymbol{v}; \beta)p(\boldsymbol{v})p(\boldsymbol{s})p(\boldsymbol{y}|\boldsymbol{i})p(\boldsymbol{i}|\boldsymbol{w}, \boldsymbol{r}, \boldsymbol{s}; \theta).$$

To obtain the likelihood expression of the observed data we need to marginalize out the latent unobserved variables, $\boldsymbol{i}, \boldsymbol{r}, \boldsymbol{w}$:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{s}, \boldsymbol{y}; \alpha, \beta, \theta) &= \sum_{\boldsymbol{i},\boldsymbol{r},\boldsymbol{w}} p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{v}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{i}, \boldsymbol{y}; \alpha, \beta, \theta) \\ &= \sum_{\boldsymbol{w}} \sum_{\boldsymbol{r}} \sum_{\boldsymbol{i}} p(\boldsymbol{y}|\boldsymbol{i})p(\boldsymbol{i}|\boldsymbol{w}, \boldsymbol{r}, \boldsymbol{s}; \theta)p(\boldsymbol{w}|\boldsymbol{x}; \alpha) \\ &\quad \times p(\boldsymbol{x})p(\boldsymbol{r}|\boldsymbol{v}; \beta)p(\boldsymbol{v})p(\boldsymbol{s}). \end{aligned}$$

Remember that the relationship between $\boldsymbol{i}^c$ and y_c is deterministic (see Eq. (4)). If we look at $\boldsymbol{i}^c$ as a binary vector, given the observed real value y_c, there are only a few *valid* vectors $\boldsymbol{i}^c$. Let $\boldsymbol{i}^c$ be a *valid* vector that assigns value $i_e^c = 0$ to all non-transferred embryos ($s_e^c = 0$) and assigns value $i_e^c = 1$ exactly y_c times among transferred embryos so that Eq. (4) is satisfied. E.g., consider a single cycle with 5 embryos, where the second and third ones were transferred ($s_2^c = s_3^c = 1$) but only one of them was implanted ($y_c = 1$): there are only 2 valid vectors, $[0, \mathbf{1}, \mathbf{0}, 0, 0]$ and $[0, \mathbf{0}, \mathbf{1}, 0, 0]$. Let $\mathcal{I}_{\boldsymbol{s},\boldsymbol{y}}$ be the set of valid vectors that assign value to all the embryos (implanted or not) according to the known outcomes $\{y_c\}_{c=1}^N$ and the selections $\{\boldsymbol{s}_:^c\}_{c=1}^N$, and $\mathcal{I}_{\boldsymbol{s}_:^c,y_c}$ the same for a specific cycle c. By the deterministic relationship, for any vector $\boldsymbol{i} \notin \mathcal{I}_{\boldsymbol{s}_:^c,y_c}$ then $p(y_c \mid \boldsymbol{i}) = 0$. We can introduce this deterministic relation in the marginalization step by summing only over the valid vectors $\boldsymbol{i} \in \mathcal{I}_{\boldsymbol{s},\boldsymbol{y}}$:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{s}, \boldsymbol{y}; \alpha, \beta, \theta) &= \sum_{\boldsymbol{w}} \sum_{\boldsymbol{r}} \sum_{\boldsymbol{i} \in \mathcal{I}_{\boldsymbol{s},\boldsymbol{y}}} p(\boldsymbol{i}|\boldsymbol{w}, \boldsymbol{r}, \boldsymbol{s}; \theta)p(\boldsymbol{w}|\boldsymbol{x}; \alpha) \\ &\quad \times p(\boldsymbol{x})p(\boldsymbol{r}|\boldsymbol{v}; \beta)p(\boldsymbol{v})p(\boldsymbol{s}) \\ &= p(\boldsymbol{x})p(\boldsymbol{v})p(\boldsymbol{s}) \sum_{\boldsymbol{r}} p(\boldsymbol{r}|\boldsymbol{v}; \beta) \\ &\quad \times \sum_{\boldsymbol{i} \in \mathcal{I}_{\boldsymbol{s},\boldsymbol{y}}} \sum_{\boldsymbol{w}} p(\boldsymbol{i}|\boldsymbol{w}, \boldsymbol{r}, \boldsymbol{s}; \theta)p(\boldsymbol{w}|\boldsymbol{x}; \alpha). \end{aligned}$$

By assuming independence among instances given the parameters, we add more structure:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{v}, \boldsymbol{s}; \alpha, \beta, \theta) &= \prod_{c=1}^{N} p(v_c) \sum_{r_c} p(r_c|v_c; \beta) \sum_{\boldsymbol{i}^c \in \mathcal{I}_{\boldsymbol{s}_:^c,y_c}} \prod_{e \in E^c} p(x_e^c)p(s_e^c) \\ &\quad \times \sum_{w_e^c} p(i_e^c|w_e^c, r_c, s_e^c; \theta)p(w_e^c|x_e^c; \alpha) \end{aligned} \tag{5}$$

as well as considering Eqs. (1)–(3).

We are interested in finding the set of parameters $\langle \alpha, \beta, \theta \rangle$ that maximize the likelihood:

$$\alpha^*, \beta^*, \theta^* = \arg\max_{\alpha,\beta,\theta} p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{v}, \boldsymbol{s}; \alpha, \beta, \theta).$$

3.3. Machine learning method

In the presented model, there are latent variables ($\boldsymbol{r}$, $\boldsymbol{w}$ and $\boldsymbol{i}$) whose value is (generally) unknown, which makes the learning of the model parameters $\langle \alpha, \beta, \theta \rangle$ difficult. We use an Expectation–Maximization (EM) algorithm [39] to overcome this issue. The EM is an iterative strategy to find (local) maximum likelihood estimators of the model parameters in the presence of missing data or latent variables. First, the expected value of the missing data is obtained. Then, the MLE parameters are obtained for that completed data.

Formally, let X be the observed variables in the model and Z the unobserved latent ones. The complete log-likelihood is $l(\eta; X, Z)$,

Table 2
Notation employed in this paper.

Symbol	Description
c	Index for cycle
e	Index for embryo
C	Set of cycles
N	Number of cycles
E^c	Set of embryos associated to cycle c
S^c	Set of embryos selected for transfer in cycle c
s_e^c	Indicates if an embryo e is selected to transfer in cycle c
$\boldsymbol{x}_e$	Characteristics of embryo e
$\boldsymbol{v}_c$	Characteristics of cycle c
w_e	Boolean random variable that represents whether embryo e is willing to implant
r_c	Boolean random variable that represents the fertility potential of patient/cycle c
i_e^c	Boolean random variable that represents whether embryo e is willing to implant in cycle c
y_c	Number of embryos implanted in cycle c

where η are the parameters which we want to estimate maximizing the likelihood.

The expectation (E) step consists in computing the conditional expected value of the log-likelihood given the observed variables and the current fit of the parameters $\eta^{(t)}$:

$$Q(\eta;\eta^{(t)}) = \mathbb{E}_{Z\sim p(z|X;\eta^{(t)})}[l(\eta;X,Z)] = \int l(\eta;X,z)p(z|X;\eta^{(t)})dz$$

where $p(z|X;\eta^{(t)})$ is the conditional probability distribution of the unobserved variables Z conditioned to the observed variables X and the current fit of the parameters $\eta^{(t)}$.

The maximization (M) step consists in finding the parameters η that maximize the conditional expectation of the E-step,

$$\eta^{(t+1)} := \operatorname*{argmax}_{\eta} Q(\eta;\eta^{(t)}).$$

In our case, the latent variables are $Z = (\boldsymbol{r}, \boldsymbol{w}, \boldsymbol{i})$, the observed ones are $X = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{v}, \boldsymbol{s})$, and the parameters $\eta = \langle \alpha, \beta, \theta \rangle$. From Eq. (5), the expected value of r_c can be probabilistically calculated in the E-step as,

$$q(r_c = r) \propto \Big(\sum_{i^c \in \mathcal{I}_{s_{:}^c, y_c}} \prod_e \sum_{w_e^c} p(i_e^c|w_e^c, r_c = r, s_e^c;\theta) p(w_e^c|\boldsymbol{x}_e^c;\alpha) \Big) p(r_c = r|\boldsymbol{v}_c;\beta) \tag{6}$$

for $r \in \{0,1\}$, where $\sum_{r\in\{0,1\}} q(r_c = r) = 1$.

Similarly, for each embryo e in cycle c, the expected value of the variable indicating embryo viability, w_e^c, is calculated as,

$$q(w_e^c = w) \propto \sum_{r_c} \Bigg(\sum_{i^c \in \mathcal{I}_{s_{:}^c, y_c}} p(i_e^c|w_e^c = w, r_c, s_e^c;\theta) p(w_e^c = w|\boldsymbol{x}_e^c;\alpha) \times \prod_{e' \neq e} \sum_{w_{e'}^c} p(i_{e'}^c|w_{e'}^c, r_c, s_{e'}^c;\theta) p(w_{e'}^c|\boldsymbol{x}_{e'}^c;\alpha) \Bigg) \times p(r_c|\boldsymbol{v}_c;\beta) \tag{7}$$

for $w \in \{0,1\}$, where $\sum_{w\in\{0,1\}} q(w_e^c = w) = 1$.

Finally, the expected value associated with each possible implantation vector, $\boldsymbol{i}$, is calculated as,

$$q(\boldsymbol{i}^c = \boldsymbol{i}) \propto \sum_{r_c} \Big(\prod_e \sum_{w_e^c} p(i_e|w_e^c, r_c, s_e^c;\theta) p(w_e^c|\boldsymbol{x}_e^c;\alpha) \Big) p(r_c|\boldsymbol{v}_c;\beta) \tag{8}$$

for all $\boldsymbol{i} \in \mathcal{I}_{s_{:}^c, y_c}$, where $\sum_{i\in\mathcal{I}_{s_{:}^c, y_c}} q(\boldsymbol{i}^c = \boldsymbol{i}) = 1$.

Then, our specific **M-step** can be expressed as finding the set of parameters $\langle \alpha, \beta, \theta \rangle$ as,

$$\operatorname{argmax}_{\alpha,\beta,\theta} \mathbb{E}_{(\boldsymbol{w},\boldsymbol{r},\boldsymbol{i})\sim q} \log p(\boldsymbol{r}, \boldsymbol{w}, \boldsymbol{i}, \boldsymbol{y}|\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{s}; \alpha, \beta, \theta)$$

where q denotes the expected values described by Eqs. (6) to (8). The conditional expectation to maximize has the following form:

$$\sum_c \sum_{i^c \in \mathcal{I}_{s_{:}^c, y_c}} q(\boldsymbol{i}^c) \Bigg[\sum_{r_c} q(r_c) \Big[\log p(r_c|\boldsymbol{v}_c;\beta) + \sum_e \sum_{w_e^c} q(w_e^c) \big[\log p(i_e^c|w_e^c, r_c, s_e^c;\theta) + \log p(w_e^c|\boldsymbol{x}_e^c;\alpha) \big] \Big] \Bigg].$$

Algorithm 1 Our EM algorithm

```
procedure EM
    t ← 0
    α^(t), β^(t), θ^(t) ← initialization()
    while q not converged do
        q ← p(i, w, r|y, x, v; α^(t), β^(t), θ^(t))      ▷ Update q: E-step (Eqs. (6), (7), (8))
        α^(t+1) ← argmax_α E_{w~q} log p(w|x; α^(t))     ▷ Update α: M1-step
        β^(t+1) ← argmax_β E_{r~q} log p(r|v; β^(t))     ▷ Update β: M2-step
        θ^(t+1) ← argmax_θ E_{i~q} log p(y|...; θ^(t))   ▷ Update θ: M3-step (Eq. (9))
        t ← t + 1
    end while
    return ⟨α, β, θ⟩
end procedure
```

The θ_1 value that maximizes this expression, i.e., the maximum likelihood estimator of θ_1, is:

$$\hat{\theta}_1 = \frac{\sum_c \sum_{i^c \in \mathcal{I}_{s_{:}^c, y_c}} \sum_e q(\boldsymbol{i}^{c'}) q(r_c = 1) q(w_e^c = 1) i_e^c}{\sum_c \sum_{i^c \in \mathcal{I}_{s_{:}^c, y_c}} \sum_e q(\boldsymbol{i}^{c'}) q(r_c = 1) q(w_e^c = 1)} \tag{9}$$

which is the probability that a viable embryo selected for transfer in a fertile cycle gets implanted. And it can be understood as our ability to model the uncertainty of the problem: the higher this probability is, the more portion of the uncertainty is modeled by the classifiers of embryos and cycles, and thus the more explanatory the model can be for new cycles. The full derivation of this expression is given in Appendix B.

As aforementioned, both Eqs. (1) and (2) are approximated by means of probabilistic classifiers, and the model parameters α and β represent the hyperparameters of the respective classifier. In this sense, the values of α and β that maximize the previous conditional expectation are obtained by learning a new fit for the classifiers using Eqs. (6) and (7), respectively, to weigh the instances of the training set.

To sum up, after initialization, our EM algorithm repeats iteratively these two steps:

(i) **Expectation:** The expectation of the latent variables r_c, w_e^c and i_e^c is computed with Eqs. (6) to (8), using the current fit of the model $\langle \alpha^{(t)}, \beta^{(t)}, \theta^{(t)} \rangle$. Note that there exist cases where we do know the value of the latent variables, r_c, w_e^c, i_e^c. When successful cycles ended up in a pregnancy ($y_c \geq 1$), we do know that the cycle was viable ($r_c = 1$), so we can safely use $q(r_c = 1) = 1$ and $q(r_c = 0) = 0$. Moreover, when the number of implanted embryos is the same as the number of transferred embryos ($y_c = |S^c|$, success rate = 1), we do know that all the transferred embryos were viable ($w_e^c = 1$), so we can safely use $q(w_e^c = 1) = 1$ and $q(w_e^c = 0) = 0$, for all $e \in S^c$. In this case, there also

exists a single valid implantation vector $\boldsymbol{i} \in \mathcal{I}_{s^c_{:},y_c}$ ($|\mathcal{I}_{s^c_{:},y_c}| = 1$), and thus $q(\boldsymbol{i}^c = \boldsymbol{i}) = 1$.

(ii) **Maximization:** A new fit of the parameters of the model $\langle \alpha^{(t+1)}, \beta^{(t+1)}, \theta^{(t+1)} \rangle$ is obtained. The probabilistic classifiers for Eq. (1) (α) and (2) (β) are learned with the weighted samples from the previous E-step. Similarly, the MLE for θ_1 is obtained with Eq. (9).

The method iterates these two steps until the stopping condition is met. The pseudocode of the resulting method is shown in Algorithm 1.

3.3.1. Set up

To initialize Algorithm 1, we assign initial probabilities directly to the sample weights (q_r for cycles, q_w for embryos, and q_i for implantation vectors) and obtain a first fit of the model with them (as if it were an M-step). All the weights are randomly generated and normalized to sum up to 1. The only exceptions are the special cases previously discussed where we actually know the value of these latent variables, for which no random initialization is required.

We have considered a stop condition that is actually two-fold: we test convergence by comparing the sample weights assigned in consecutive iterations, and we fix a maximum number of iterations (100) that the algorithm can run.

As known, the EM strategy is only guaranteed to reach local maxima or saddle points of the likelihood. We run our algorithm multiple (10) times with different initializations to try to reach other local maxima and keep only the best one, mitigating thus the local-maximum problem of EM algorithms.

4. Empirical validation

In this section, we aim to perform a robust validation of the proposed model. To do so, we use different probabilistic classifiers for our embryo viability (Eq. (1)) and cycle fertility potential (Eq. (2)) modules, and compare them to others learned with up to 3 different baseline approaches. In particular, we carry out three sets of experiments:

- Experiment #1: we compare the results of our model against the classifiers obtained with a series of baseline approaches to test the behavior of our proposal.
- Experiment #2: we estimate the relevance of the information of the cycle in the embryo implantation predictive task by including the cycle's characteristics as descriptive variables for the baseline approaches too.
- Experiment #3: we validate our model by comparing its results to the ASEBIR protocol. Specifically, we compare the performance of our model using or not this score as a feature.

The interpretation of the predictions that we can obtain from our model needs proper consideration. For instance, by using the whole model we obtain the probability of implantation of an embryo in a cycle (it assumes independence between embryo and cycle), which is calculated as:

$$p(i^c_e = 1 | x^c_e, s^c_e, v_c; \alpha, \beta, \theta) = p(i^c_e = 1 | w^c_e = 1, s^c_e, r_c = 1; \theta) \times p(w^c_e = 1 | x^c_e; \alpha) p(r_c = 1 | v_c; \beta) \quad (10)$$

where $p(i^c_e = 1 | w^c_e = 1, s^c_e, r_c = 1; \theta) = \theta_1 \cdot s^c_e$. Remember that if $s^c_e = 0$, $p(i^c_e = 1 | w^c_e, s^c_e = 0, r_c; \theta) = 0$. This is the reason why the evaluation will be only performed with embryos that were transferred ($s^c_e = 1$). The other two terms (Eqs. (1) and (2)) represent the probabilistic classifiers of embryo and cycle viability, respectively. In fact, we could unplug these classifiers from the learned model and use them to predict embryo/cycle viability. Note that in real practice, in embryo-selection time, the cycle's stimulation is already finished; thus, if we fix the cycle, the ranking of embryos given by Eqs. (1) and (10) is the same.

As aforementioned, we approximate Eqs. (1) and (2) by means of probabilistic classifiers. By the no free lunch theorem, we know that different classifiers may perform differently depending on the context. In order to make a fair comparison, in these experiments we have tested three types of classifiers of different nature: Logistic Regression (LR), Random Forest (RF) and Gradient Boosting (GBOOST) classifiers. We use the default parametrization of these techniques, as implemented by Python's Scikit-learn library [40].

4.1. Baseline approaches

The two main characteristics of our model are the way it combines the information from the cycle and the embryos, as well as its ability to learn using all the available examples independently of the amount of class information that they carry. The baselines that we use in this study for comparison follow simplistic approaches to these two aspects. All of them use the same types of probabilistic classifiers previously described, for the sake of fair comparison. Note that the models learned with these baseline methods directly predict implantation. This is slightly different from the embryo module of our probabilistic model (Eq. (1)), which actually predicts whether an embryo is willing to implant (viability).

Baseline approaches use, to learn the classifiers, a transformed dataset using different assumptions to assign a label to examples that are originally (partially) unlabeled (Table 1 summarizes the number of embryos in our database with (un)known fate). We design the baseline approaches in an incremental way regarding these assumptions on the partially labeled embryos, as summarized in Table 3.

Our *pessimistic* approach assumes that all the embryos with unknown fate are negative examples (unviable embryos). This is the simplest decision as it leads to a completely labeled dataset that can be directly learned using standard supervised learning techniques. However, it holds a heavy assumption: all non-transferred embryos are not viable for implantation (questionable), and all the embryos in partially implanted cycles are not viable for implantation (wrong: some of them are, but we do not know their identity). This approach brings a severe class imbalance problem, with only a tiny portion of embryos labeled as positive.

With the objective of relaxing this heavy assumption, our second baseline approach does not assign any label to embryos of unknown fate (see Table 3). This decision leads to a semi-supervised learning setting. We use a standard EM algorithm [41], which we call *simple EM*, to learn from this type of data. Thus, we allow the learning technique to unveil the class label of the embryos of unknown fate, alleviating at the same time the class imbalance problem.

This previous approach still dismisses the class information of embryos in partially implanted cycles: the label proportions or counts of implanted embryos per cycle. Our third baseline approach lets the model be learned from these counts of implanted embryos per cycle (see Table 3). This decision leads to a learning from label proportions setting. We use the EM algorithm proposed by Hernández-González et al. [8], which we call *LP-EM*, to learn from this type of data. One can arguably consider that this approach uses all the available information of supervision.

4.2. Evaluation

As a weakly supervised problem [42], a fair evaluation of the models is not trivial and needs to be properly addressed. Fully unlabeled examples (embryos non-transferred) might be used for learning but not for model performance assessment. Fortunately, partially labeled examples (transferred embryos in cycles with partial implantation), where only the proportion of implanted embryos is known, can carefully be used for evaluation.

Performance is assessed in terms of different metrics, which are applied in each of the experiments only if all the required information is available.

Table 3
Description of the labeling used by the baselines. In each method, embryos of different (un)known fates receive these labels: negative (0), positive (1), unknown (?), or label-proportions (lp).

Method	Failed	Partial implantation	Implanted	Non-transferred
Pessimistic	0	0	1	0
Simple EM	0	?	1	?
LP-EM	0	lp	1	?

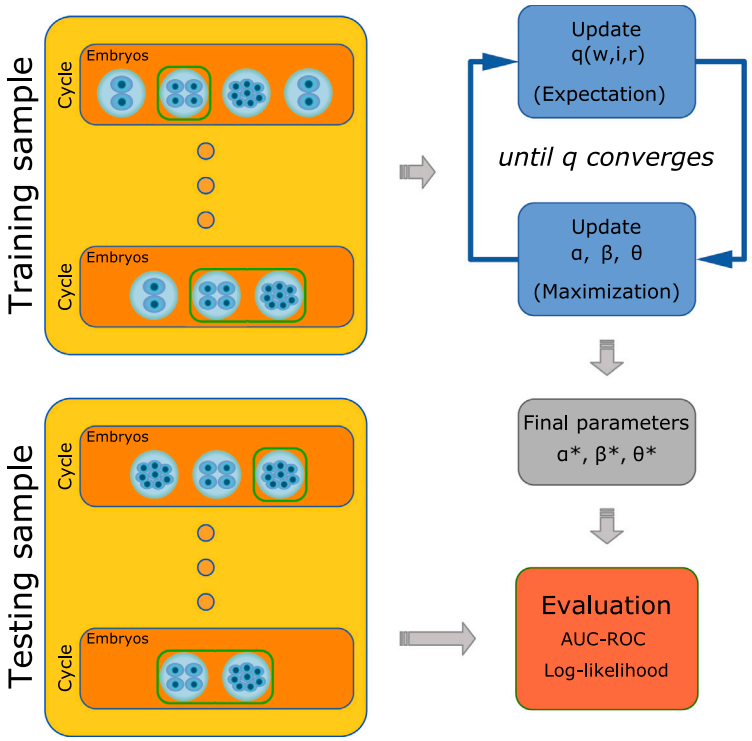


Fig. 4. Workflow of our study.

To test the ability to predict embryo implantation, we use the area under the ROC curve (AUC–ROC) [43,44], which plots the true positive rate against the false positive rate as the discrimination threshold is varied. It does not require fixing a threshold to estimate the predictive performance of a probabilistic classifier. It represents the probability that the classifier will assign to a positive instance higher probability than to a negative one. The higher the score, the better the classifier (a random classifier would obtain a value of 0.5). Note that AUC–ROC could not be appropriate when the dataset is highly unbalanced, as it produces overly optimistic measurements [45,46]. Many alternatives have been proposed to address this limitation, including partial AUC–ROC [46] (which focuses on the most relevant parts of the ROC curves) and the area under the Precision–Recall curve [45] (which focuses exclusively on the minority and relevant class). In our study, this issue is expected to impact mainly one of our baselines, the pessimistic approach, the results of which should be interpreted accordingly. As all the labels of the individual embryos are needed to calculate this metric, only those belonging to failed or completely-implanted cycles are considered.

To account also for the embryos in cycles with partial implantation, we use the negative log-likelihood. It measures the confidence of the model in predicting each of the labels. Formally, we calculate the mean probability of the real number of implanted embryos per cycle given the current model:

$$\mathcal{L}(\boldsymbol{Y};\alpha,\beta,\theta) = -\frac{1}{N}\sum_{c=1}^{N}\sum_{j=0}^{|S^c|}\mathbb{I}[y_c = j]\log p(y_c) \tag{11}$$

where $|S^c|$ is the number of transferred embryos in cycle c, and $p(y_c)$, the probability of cycle c having y_c implanted embryos over all valid implantation vectors $\boldsymbol{i}^c$, is,

$$p(y_c) = \sum_{\boldsymbol{i}^c \in \mathcal{I}_{s^c_{:},y_c}} \prod_{e} \left[i^c_e p(i^c_e = 1) + (1 - i^c_e) p(i^c_e = 0) \right] \tag{12}$$

where $p(i^c_e)$ is given by Eq. (10).

For model performance assessment, we use 10 × 5-fold cross-validation. All the results show the average value. Fig. 4 displays the workflow of this study.

4.3. Experiment #1: Performance comparison

In this first snapshot of the experiments, we show a comparison between our model and the different baseline approaches when used for embryo implantation prediction, for different base probabilistic classifiers. Table 4 shows the results in terms of different metrics.

Our model obtains the best performance in terms of the AUC–ROC metric (consistently for all the classifier types). From the detailed inspection of the densities produced by these approaches,[1] we can appreciate signs of learning, though they might be limited, for classifiers learned with all the approaches (e.g., the high-quality embryos receive a higher probability of implantation). However, the results of the baselines in terms of AUC–ROC are rather limited. As mentioned previously, AUC–ROC is calculated using only embryos with known fate. Thus, it is reasonable to think that it favors those approaches which can detect if the cycle is actually a critical factor. In this set of

[1] Figures available in the supplementary material at https://jhernandezgonzalez.github.io/supp_arts_pgm.html

Table 4
Results in terms of AUC–ROC and log-likelihood of classifiers of different type learned with our model and the 3 baseline approaches.

Method	Base classifier	AUC–ROC	Log-likelihood
Pessimistic	LR	0.58 ± 0.06	0.63 ± 0.18
Simple EM		0.56 ± 0.06	0.51 ± 0.11
LP-EM		0.56 ± 0.06	0.47 ± 0.05
Our model		0.62 ± 0.08	0.52 ± 0.10
Pessimistic	RF	0.61 ± 0.06	0.62 ± 0.18
Simple EM		0.55 ± 0.07	0.46 ± 0.11
LP-EM		0.58 ± 0.08	**0.42 ± 0.06**
Our model		0.71 ± 0.05	**0.42 ± 0.07**
Pessimistic	GBOOST	0.62 ± 0.05	0.67 ± 0.26
Simple EM		0.61 ± 0.08	0.51 ± 0.12
LP-EM		0.60 ± 0.08	0.44 ± 0.05
Our model		**0.73 ± 0.07**	0.43 ± 0.06

Table 5
Results in terms of AUC–ROC and log-likelihood of classifiers of different type learned with our model and the *pessimistic* baseline (considering or not the cycle features).

Method	Base classifier	AUC–ROC	Log-likelihood
Pessimistic	LR	0.58 ± 0.06	0.63 ± 0.18
Pessimistic with cycle feat.		0.63 ± 0.07	0.70 ± 0.25
Our model		0.62 ± 0.08	0.52 ± 0.10
Pessimistic	RF	0.61 ± 0.06	0.62 ± 0.18
Pessimistic with cycle feat.		**0.74 ± 0.06**	0.58 ± 0.15
Our model		0.71 ± 0.05	**0.42 ± 0.07**
Pessimistic	GBOOST	0.62 ± 0.05	0.67 ± 0.26
Pessimistic with cycle feat.		0.72 ± 0.05	0.64 ± 0.12
Our model		0.73 ± 0.07	0.43 ± 0.06

experiments, the only approach that uses the cycle information is our complete model. This could provide an explanation for the performance gap observed in the results, and it is precisely the idea that we test in the second set of experiments.

In terms of negative log-likelihood, which measures the confidence of the model about its predictions and allows us to use also the partially implanted cycles for evaluation, the *pessimistic* approach, which deals with a highly unbalanced dataset due to its unrealistic but simplifying assumption, shows the worst results. The EM-based approaches perform better: they use partially labeled cycles in model learning without any hard assumptions. These approaches, mainly the one that considers also label proportions, seem promising as their performance reaches that of our model: they match the best global result (with RF classifiers) of our model and even outperform it when learning LR classifiers.

4.4. Experiment #2: cycle characteristics for embryo implantation prediction

In this second snapshot of the experiments, we pay attention to a different dimension of the problem: the importance of the features describing the cycle configuration for the predictive task of embryo implantation. Our model includes them, whereas the baselines do not (as configured so far). Table 5 shows the results in terms of different metrics for our model and the pessimistic approach (already shown in the previous section), together with the results of the pessimistic approach when the training dataset is enlarged with the cycle features.

The results in terms of log-likelihood are not conclusive: in some cases, the inclusion of cycle features improves the performance of the models, but it is not consistent. However, we can observe clearly a relevant improvement in the results in terms of AUC–ROC, competitive with our complete model. Inspecting the density of the probability values (available in the Supplementary Material), we observe that, although true unviable embryos are clearly concentrated around probability equal to 0, the density for the truly implanted embryos is clearly shifted towards higher probability values. That is, in some cases where the cycle is identified as viable, the models are more confident when predicting implantation. This seems to explain the differences in the AUC–ROC values of the pessimistic approach using or not the cycle features.

Table 6
Results in terms of AUC–ROC and log-likelihood of classifiers of different type learned with our model (considering or not the ASEBIR score as a feature). The last column shows the mean value learned for the θ_1 model parameter.

Model version	Base classifier	AUC–ROC	Log-likelihood	θ_1
With ASEBIR score	LR	0.63 ± 0.08	0.51 ± 0.10	0.52 ± 0.01
Without ASEBIR score		0.62 ± 0.08	0.52 ± 0.10	0.51 ± 0.00
With ASEBIR score	RF	**0.71 ± 0.05**	**0.42 ± 0.07**	0.48 ± 0.01
Without ASEBIR score		0.71 ± 0.05	**0.42 ± 0.07**	0.48 ± 0.01
With ASEBIR score	GBOOST	**0.71 ± 0.04**	0.45 ± 0.05	0.49 ± 0.00
Without ASEBIR score		**0.73 ± 0.07**	0.43 ± 0.06	0.49 ± 0.01

To fully grasp the behavior of our model, we inspect the probability densities for successful and failed cycles in Fig. 5, separately for embryo viability prediction (Eq. (1), left column), cycle fertility-potential prediction (Eq. (2), middle column), and cycle success prediction (whole model, right column). An ideal model would completely separate the densities in this last column. The results of all classifiers show a large intersection between both densities, but the mode of the density for successful cycles (pregnancy) is shifted to the right of the density of the failed cycles. This points out a small signal: the model seems to predict success, on average, more for actually implanted embryos than for those that failed. According to the plots of the first column (embryo viability), there is almost no difference between successful and failed treatments. At a first glance, embryos seem to be irrelevant to predict a pregnancy. Nevertheless, it is noteworthy that the embryos employed in this part of the study are only the transferred ones, that is, a subset of the set of embryos manually selected by the embryologists as the best embryos for transference (see Fig. 2(a)). Most of the predictive power of the model seems to come from the **cycle** descriptors. The middle column of Fig. 5 shows that cycles that actually induced a pregnancy receive a higher probability of cycle viability. One can conclude that the protocol followed by the embryologists for embryo selection based on the morphological features performs well, as our model seems not to be able to further discriminate the embryos based on this data (the same they used) alone.

4.5. Experiment #3: Our model and the effect of the ASEBIR score

So far, we have focused on analyzing the performance of our model and the baselines regarding their ability to predict ART success. We have also available a measure of embryo quality, calculated by the embryologists according to the ASEBIR protocol [1], for each individual embryo in our database. In this last set of experiments, we test whether our model agrees with the ASEBIR quality score.

To study the agreement between our method and the ASEBIR score, we compare two versions of our complete model: (i) a model trained with an embryos dataset where the ASEBIR score is just another descriptive feature (the ASEBIR score is an element in vector $\boldsymbol{x}$), and (ii) a model trained with a dataset from which the ASEBIR score has been completely removed. In Table 6, we show the results obtained with both models (with and without the ASEBIR score feature) for the different probabilistic classifiers.

It is noteworthy that there are no significant differences between both models. Including the ASEBIR score as a descriptive feature of the embryos does not apparently boost the performance of the model. Table 6 also shows the mean value estimated for the θ_1 parameter, which measures the probability that a viable embryo actually gets implanted in a viable cycle. It represents the third source of uncertainty in our proposal, which measures the effect of any unknown factors. Its value is usually close to 0.5. This means that in these cases, even if the classifiers consider that both embryo and cycle are viable, the

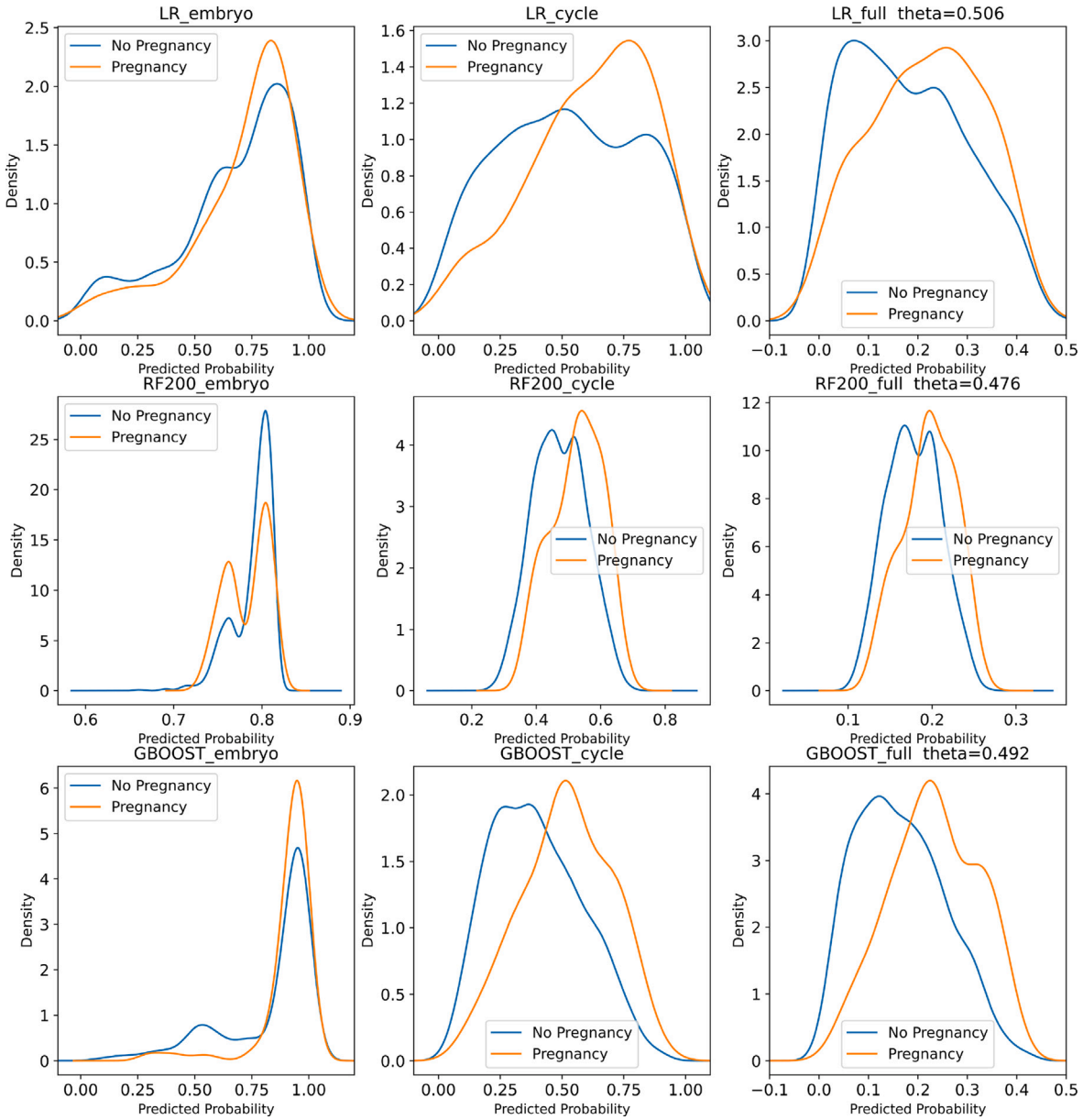


Fig. 5. Densities of the predicted probabilities of our model separated by outcome (pregnancy or not). Each row shows results with different types of base classifiers. Each column shows densities of the predicted probabilities (i) for embryo viability (left column), (ii) for cycle viability (middle column), and (iii) for cycle success (whole model, right column).

model expects that only half of these pairs will succeed. The standard deviation is low, implying a consistent estimation.

As before, we also inspect the probability densities to understand the behavior of the model regarding the ASEBIR score in Fig. 6. Specifically, we show the results with the version of the model that **does not use** the ASEBIR score feature. Under the independence hypothesis, the quality of an embryo should not be related to the probability that a cycle is viable and, in general, we observe that the embryo information has not leaked into the cycle classifier (the probability density of cycle viability –middle column– is almost the same for all the ASEBIR categories). Embryo quality has the highest impact on the model ability to predict embryo viability (left column). All classifiers (mainly LR and GBOOST) tend to separate the best (A) and worst (D) quality embryos, but barely discriminate embryos of medium quality (B and C). The model mostly agrees with the ASEBIR score in the identification of the most and least promising embryos, even without explicitly considering the ASEBIR score as a feature.

All in all, although our method seems not to directly consider the ASEBIR score feature as relevant, it is important to bear in mind that the rest of descriptive variables of the embryos, $\boldsymbol{x}$, are exactly the ones used in ASEBIR protocol [1]. Given the alignment between the ASEBIR score and our model's results observed in Fig. 6, the irrelevance of the ASEBIR score feature is possibly due to the fact that our method finds the relevant information among the rest of variables when this key feature is not given. This interpretation would suggest that the ASEBIR protocol already extracts the relevant information out of the available morphological features, which is also captured by our model.

5. Conclusions

In this work, we address the problem of embryo selection for ARTs, a complex real-world problem with partial observability issues. We propose a novel probabilistic graphical model, an extension of the standard embryo–uterine model, which assumes independence between embryos and cycles. It is, to the extent of our knowledge, the first one that takes into account three different possible sources of uncertainty, accounting for the unknown factors which cause that viable embryos, selected

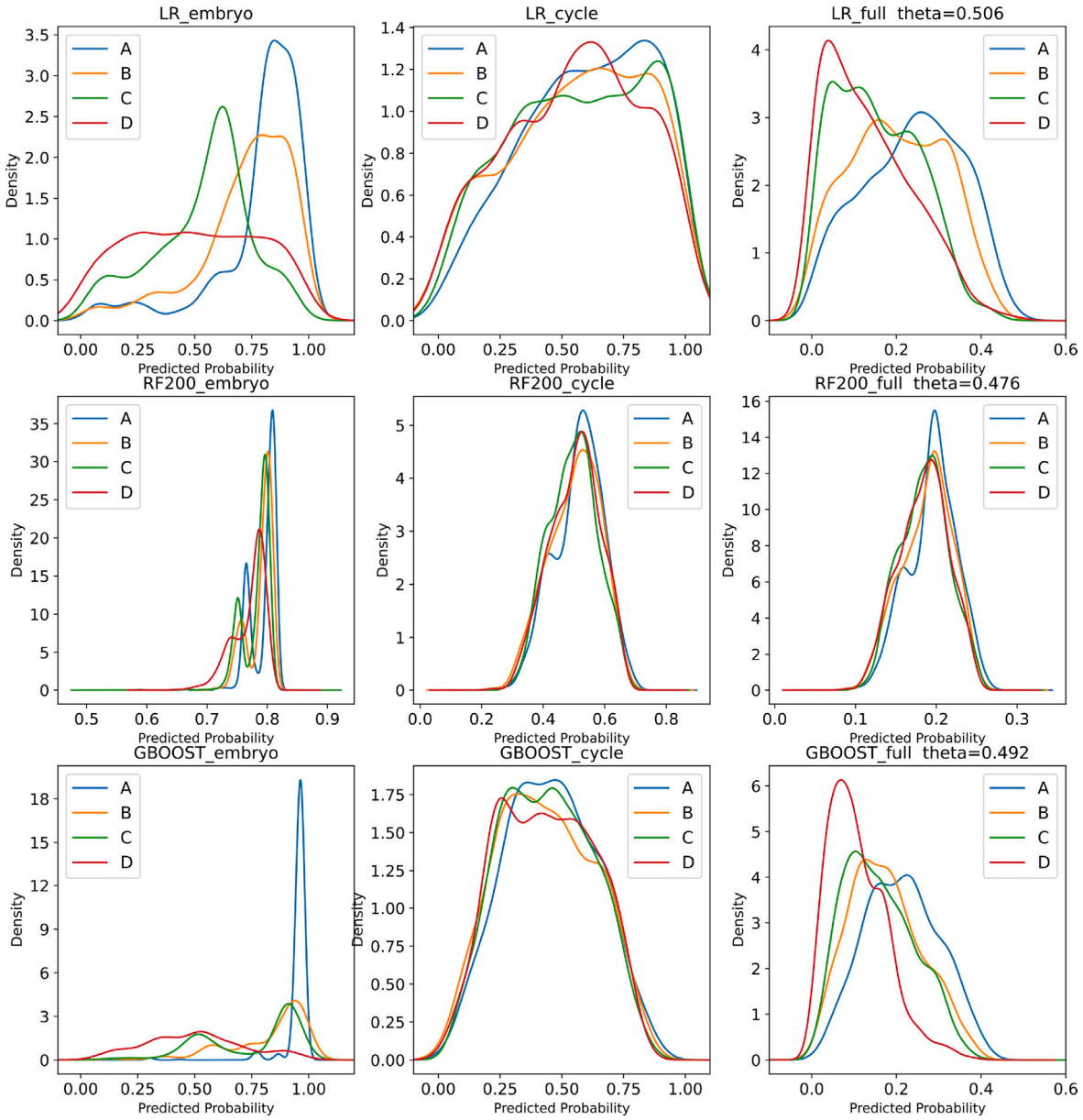


Fig. 6. Densities of the predicted probabilities of our model separated by ASEBIR quality category. The learned models **do not use** the ASEBIR quality score as a descriptive feature. Each row shows results with different types of base classifiers. Each column shows densities of the predicted probabilities (i) for embryo viability (left column), (ii) for cycle viability (middle column), and (iii) for cycle success (whole model, right column).

to transfer, fail to implant. We also derived its learning procedure, which is able to learn from all the available information of supervision, including partially labeled data. Using morphological data for each individual embryo and characteristics of the cycle, the model is able to predict embryo implantation.

We studied the effect of the ASEBIR embryo's quality score within our model. The models learned with and without the ASEBIR score show a similar separation between categories. Our results suggest that, once embryologists have made their selection, the model does not provide more information about individual embryos. This might indicate that the current protocol already extracts most of the value out of the available morphological data. The performance of the model was further validated against three baseline approaches. We show the benefits of implementing an EM strategy for the learning process, letting the learning technique unveil the label of embryos of unknown fate. We observe that the cycle's features play a key role to predict implantation, especially when either all embryos in a cycle or none were implanted. More importantly, we obtain an estimation of the uncertainty originating from unknown, external factors, θ_1. The most common result suggests that even when the embryo and cycle are viable, there is only about a 50% probability of actually inducing pregnancy. The novel result increases the modeling ability of the system and may assist clinicians in decision-making in real ART practice.

Many issues are still open. The empirical validation of the method by means of an enlarged experimental setting is still possible, as well as using real data from more than a hospital/source. Moreover, the learning techniques of the classifiers could be fine-tuned to optimize their predictive power, as we only used default configurations. Another direction would be to conceive new, maybe simpler, PGMs to test the assumptions of our current model (independence between embryos and cycles, awareness of a third source of error, etc.). Finally, the most challenging idea for future work would be to try to validate, in collaboration with embryologists, the value for θ_1 obtained by our model and its relationship with the proportion of promising treatments that failed to implant.

Table A.7
Features collected for each ART cycle.

Variable	Possible values	Description
CycleId	*Numeric*	Identifier of the cycle
TEsteril	*Numeric*	Time since infertility was detected
Indicac	endometriosi, iafailed, tubal, male, mix, unknown	Indication of the cycle
Features related to female patient		
Age	*Numeric*	Age
BMI	*Numeric*	Body mass index
PRegPrev	No, Yes	Has she ever got pregnant?
AboPrev	No, Yes	Has she ever aborted?
FSH	*Numeric*	Quantity of follicle-stimulating hormone
CiclesPrev	*Numeric*	Number of previously undergone cycles
AMH	*Numeric*	Quantity of anti-mullerian hormone
folAntral	*Numeric*	Number of antral follicles
E2	*Numeric*	Quantity of estradiol
P4	*Numeric*	Quantity of progesterone
lEnd	*Numeric*	Endometrial thickness
Features related to male patient		
qaSemen	A, N, O, OA, OAT	Quality of the semen
REM	*Numeric*	Total progressive sperm recovery
Features related to stimulation		
Protocol	PC, PL	Stimulation protocol
Stimul	FSH+Lhrec, FSHrec, FSHrec+hMG, FSHur, FSHur+hMG, hMG	Stimulation treatment
dEst	*Numeric*	Number of days of stimulation
unidFSH	*Numeric*	Units of FSH
unidLH	*Numeric*	Units of LH
Summary of embryos		
nObtenEmb	*Numeric*	Number of embryos finally obtained
FertilRate	*Numeric*	nObtenEmb/Number of mature oocytes (MII state)
nTransfEmb	*Numeric*	Number of transferred embryos
SuccessRate	*Numeric*	Number of implanted embryos/nTransfEmb

CRediT authorship contribution statement

Jerónimo Hernández-González: Conceptualization, Methodology, Implementation, Supervision, Writing. **Olga Valls:** Implementation, Methodology, Writing. **Adrián Torres-Martín:** Implementation, Methodology, Writing. **Jesús Cerquides:** Conceptualization, Methodology, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

J.H.-G. is a Serra Húnter Fellow. This work was partially supported by projects Crowd4SDG (No 872944) and Humane-AI-net (No 952026), funded by EU Horizon 2020 program, and project CI-SUSTAIN (PID2019-104156GB-I00), funded by Spanish Ministry of Science and Innovation, Spain. We would like to thank our collaborator embryologists: the ART unit from Hospital Donostia (Spain) who collected the data and granted us permission to use it, and Núria Correa (Eugin clinic) who provided meaningful advice and checked the manuscript.

Appendix A. Feature descriptions

The features collected for each ART cycle are shown in Table A.7.

The features collected for each embryo are shown in Table A.8.

Appendix B. Complete derivation of the model parameter estimation

The update of the model parameters α, β, θ (M-step) can be expressed as

$$\arg\max_{\alpha,\beta,\theta} \mathbb{E}_{(\boldsymbol{w},\boldsymbol{r},\boldsymbol{i})\sim q} \log p(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{r}, \boldsymbol{i} | \boldsymbol{x}, \boldsymbol{v}, \boldsymbol{s}; \alpha, \beta, \theta)$$

Let us imagine that we know the real value of all hidden variables. Thus, the likelihood would be

$$\prod_c \prod_{\boldsymbol{i}^{c'}} \left[\prod_{r'_c} \left[p(r'_c | v_c; \beta) \times \prod_e \prod_{w_e^{c'}} \left[p(i_e^{c'} | w_e^{c'}, \boldsymbol{r}'_c, s_e^c; \theta) p(w_e^{c'} | x_e^c; \alpha) \right]^{\mathbb{I}[w_e^{c'} = w_e^c]} \right]^{\mathbb{I}[r'_c = r_c]} \right]^{\mathbb{I}[\boldsymbol{i}^{c'} = \boldsymbol{i}^c]}$$

and the log-likelihood:

$$\sum_c \sum_{\boldsymbol{i}^{c'}} \mathbb{I}[\boldsymbol{i}^{c'} = \boldsymbol{i}^c] \left[\sum_{r'_c} \mathbb{I}[r'_c = r_c] \left[\log p(r'_c | v_c; \beta) + \sum_e \sum_{w_e^{c'}} \mathbb{I}[w_e^{c'} = w_e^c] \left[\log p(i_e^{c'} | w_e^{c'}, \boldsymbol{r}'_c, s_e^c; \theta) + \log p(w_e^{c'} | x_e^c; \alpha) \right] \right] \right]$$

But, if the real values are unknown, we need to resort to the expected values as,

$$\sum_c \sum_{\boldsymbol{i}^{c'} \in \mathcal{I}_{\boldsymbol{s}^c_\cdot, \boldsymbol{y}_c}} q(\boldsymbol{i}^{c'}) \left[\sum_{r'_c} q(r'_c) \left[\log p(r'_c | v_c; \beta) + \sum_e \sum_{w_e^{c'}} q(w_e^{c'}) \left[\log p(i_e^{c'} | w_e^{c'}, \boldsymbol{r}'_c, s_e^c; \theta) + \log p(w_e^{c'} | x_e^c; \alpha) \right] \right] \right]$$

Note that the variables i follow a Bernoulli distribution:

$$i_e^c \sim Bernoulli(\theta_{r_c \cdot w_e^c \cdot s_e^c})$$

where, in practice, θ_0 fixed to $\theta_0 = 0$ (whenever r_c, w_e^c, or s_e^c are zero: no transfer, or bad cycle/embryo) and θ_1 determines the probability of

Table A.8
Features collected for each individual oocyte/embryo.

Variable	Possible values	Description
CycleId	*Numeric*	Identifier of the cycle
EmbryoId	*Numeric*	Identifier of the embryo
Technique	IVF, ICSI	Fertilization technique
Features related to oocyte stage		
Vac	No, Few, Many	Presence of vacuoles
Rel	No, Yes	Presence of smooth endoplasmic reticulum clusters
EPV	Normal, Augmented	Description of the perivitelline space
CP	Normal, Abnormal	Description of the first polar body
PN	*Numeric*	Tesarik and Greco's pronuclear grade
Features at day 1 after fertilization		
CP+1	*Numeric*	Number of polar bodies
Z	Z1, Z2, Z3, Z4	Scott's pronuclear grade
Features at day 2 after fertilization		
nCel+2	*Numeric*	Number of cells
frag+2	*Numeric*	Percentage of cell fragmentation
simet+2	No, Yes	Symmetry of the cells
ZP+2	Normal, Abnormal	Pellucid zone
vac+2	No, Few, Many	Presence of vacuoles
multiNuc+2	No, Yes	Presence of multi-nucleation in a cell
Quality+2	A, B, C, D	ASEBIR quality grade
Transfer	No, Yes	Embryo selected for transference

implantation in perfect conditions. To find the parameter θ_1, we derive the log-likelihood with respect to θ_1, and set it to 0:

$$\frac{\partial}{\partial\theta_1}\Big[\sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} q(\boldsymbol{i}^{c\prime}) q(r_c=1) \times \sum_e q(w^c_e=1)\big[i^{c\prime}_e \log\theta_1 + (1-i^{c\prime}_e)\log(1-\theta_1)\big]\Big] = 0$$

which reduces to,

$$\sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} \sum_e q(\boldsymbol{i}^{c\prime}) q(r_c=1) q(w^c_e=1)\left[\frac{i^{c\prime}_e}{\theta_1} - \frac{(1-i^{c\prime}_e)}{(1-\theta_1)}\right] = 0$$

and, after rearrangement,

$$\sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} \sum_e q(\boldsymbol{i}^{c\prime}) q(r_c=1) q(w^c_e=1)\left[(1-\theta_1) i^{c\prime}_e\right] = \sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} \sum_e q(\boldsymbol{i}^{c\prime}) q(r_c=1) q(w^c_e=1)\left[(1-i^{c\prime}_e)\theta_1\right]$$

then,

$$\sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} \sum_e q(\boldsymbol{i}^{c\prime}) q(r_c=1) q(w^c_e=1) i^{c\prime}_e = \sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} \sum_e q(\boldsymbol{i}^{c\prime}) q(r_c=1) q(w^c_e=1)\theta_1$$

and, thus, we reach the final formula for the update of θ_1:

$$\theta_1 = \frac{\sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} \sum_e q(\boldsymbol{i}^{c\prime}) q(r_c=1) q(w^c_e=1) i^{c\prime}_e}{\sum_c \sum_{\boldsymbol{i}^{c\prime}\in I_{s^c_{:,y_c}}} \sum_e q(\boldsymbol{i}^{c\prime}) q(r_c=1) q(w^c_e=1)}$$

References

[1] M. Ardoy, G. Caderón, G. Arroyo, J. Cuadros, M.J. Figueroa, R. Herrer, J.M. Moreno, Á. Ortiz, F. Prados, L. Rodríguez, J. Santaló, M.J. de los Santos, J. Ten, M.J. Torelló, ASEBIR criteria for the morphological evaluation of human oocytes, early embryos and blastocysts, in: Clinical Embryology Papers, second ed., II, Asociación para el Estudio de la Biología de la Reproducción (ASEBIR), Madrid, Spain, 2008, pp. 1–59.

[2] Spanish Society of Fertility, National Registry of Activity 2018 - Registry SEF, Technical Report, Spanish Ministry of Health, 2018, https://www.registrosef.com/public/docs/sef2018_IAFIVm.pdf.

[3] L. Engmann, N. Maconochie, S. Tan, J. Bekir, Trends in the incidence of births and multiple births and the factors that determine the probability of multiple birth after IVF treatment, Hum. Reprod. 16 (2001) 2598–2605, http://dx.doi.org/10.1093/humrep/16.12.2598.

[4] C.R. ESHRE Campus, Prevention of twin pregnancies after IVF/ICSI by single embryo transfer, Hum. Reprod. 16 (4) (2001) 790–800, http://dx.doi.org/10.1093/humrep/16.4.790.

[5] I. Cuevas Saiz, M.C. Pons Gatell, M. Cuadros Vargas, A. Delgado Mendive, N. Rives Enedáguila, M. Moragas Solanes, B. Carrasco Canal, J. Teruel López, A. Busquets Bonet, M.V. Hurtado de Mendoza Acosta, The Embryology Interest Group: updating ASEBIR's morphological scoring system for early embryos, morulae and blastocysts, Medicina Reproductiva Y Embriología Clínica 5 (1) (2018) 42–54, http://dx.doi.org/10.1016/j.medre.2017.11.002.

[6] G. Corani, M.C. Magli, A. Giusti, L. Gianaroli, L.M. Gambardella, A Bayesian network model for predicting pregnancy after in vitro fertilization, Comput. Biol. Med. 43 (2013) 1783–1792, http://dx.doi.org/10.1016/j.compbiomed.2013.07.035.

[7] F. Guérif, A. le Gouge, B. Giraudeau, J. Poindron, R. Bidault, O. Gasnier, D. Royère, Limited value of morphological assessment at days 1 and 2 to predict blastocyst development potential: a prospective study based on 4042 embryos, Hum. Reprod. 22 (7) (2007) 1973–1981, http://dx.doi.org/10.1093/humrep/dem100.

[8] J. Hernández-González, I. Inza, L. Crisol-Ortíz, M.A. Guembe, M.J. Iñarra, J.A. Lozano, Fitting the data from embryo implantation prediction: Learning from label proportions, Stat. Methods Med. Res. 27 (2018) 1056–1066, http://dx.doi.org/10.1177/0962280216651098.

[9] M. Kragh, J. Rimestad, J. Berntsen, H. Karstoft, Automatic grading of human blastocysts from time-lapse imaging, Comput. Biol. Med. 115 (2019) 103494, http://dx.doi.org/10.1016/j.compbiomed.2019.103494.

[10] A. Torres-Martín, J. Hernández-González, J. Cerquides, Validation on real data of an extended embryo-uterine probabilistic graphical model for embryo selection, in: Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence (CCIA'21), 2021, http://dx.doi.org/10.3233/FAIA210139.

[11] S. Seshadri, S.K. Sunkara, Natural killer cells in female infertility and recurrent miscarriage: a systematic review and meta-analysis, Hum. Reprod. Update 20 (3) (2014) 429–438, http://dx.doi.org/10.1093/humupd/dmt056.

[12] I. Gat, J. Levron, G. Yerushalmi, J. Dor, M. Brengauz, R. Orvieto, Should zygote intrafallopian transfer be offered to all patients with unexplained repeated in-vitro fertilization cycle failures? J. Ovarian Res. 7 (7) (2014) http://dx.doi.org/10.1186/1757-2215-7-7.

[13] L.T. Polanski, M.N. Baumgarten, S. Quenby, J. Brosens, B.K. Campbell, N.J. Raine-Fenning, What exactly do we mean by 'recurrent implantation failure'? A systematic review and opinion, Reproductive BioMedicine Online 28 (4) (2015) 409–423, http://dx.doi.org/10.1016/j.rbmo.2013.12.006.

[14] N. Zaninovic, Z. Rosenwaks, Artificial intelligence in human in vitro fertilization and embryology, Fertility and Sterility 114 (5) (2020) 914–920, http://dx.doi.org/10.1016/j.fertnstert.2020.09.157.

[15] C. Siristatidis, A. Pouliakis, C. Chrelias, D. Kassanos, Artificial intelligence in IVF: A need, Syst. Biol. Reproductive Medicine 57 (4) (2011) 179–185, http://dx.doi.org/10.3109/19396368.2011.558607.

[16] M. Simopoulou, K. Sfakianoudis, E. Maziotis, N. Antoniou, A. Rapani, G. Anifandis, P. Bakas, S. Bolaris, A. Pantou, K. Pantos, M. Koutsilieris, Are computational applications the "crystal ball" in the IVF laboratory? The evolution from mathematics to artificial intelligence, J. Assis. Reproduction and Genet. 35 (2018) 1545–1557, http://dx.doi.org/10.1007/s10815-018-1266-6.

[17] E.I. Fernandez, A.S. Ferreira, M.H.M. Cecílio, D.S. Chéles, R.C.M. de Souza, M.F.G. Nogueira, J.C. Rocha, Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data, J. Assist. Reproduction and Genet. 37 (10) (2020) 2359–2376, http://dx.doi.org/10.1007/s10815-020-01881-9.

[18] S.A. Roberts, Models for assisted conception data with embryo-specific covariates, Stat. Med. 26 (1) (2007) 156–170, http://dx.doi.org/10.1002/sim.2525.

[19] G. Patrizi, C. Manna, C. Moscatelli, L. Nieddu, Pattern recognition methods in human-assisted reproduction, Int. Trans. Oper. Res. 11 (2004) 365–379, http://dx.doi.org/10.1111/j.1475-3995.2004.00464.x.

[20] C. Manna, G. Patrizi, A. Rahman, H. Sallam, Experimental results on the recognition of embryos in human assisted reproduction, Reproductive Biomedicine Online 8 (4) (2004) 460–469, http://dx.doi.org/10.1016/S1472-6483(10)60931-5.

[21] M.G. Minasi, P. Greco, M.T. Varricchio, P. Barillari, E. Greco, The clinical use of time-lapse in human-assisted reproduction, Therap. Adv. Reproductive Health 14 (2020) 2633494120976921, http://dx.doi.org/10.1177/2633494120976921.

[22] D.C. Kieslinger, S. De Gheselle, C.B. Lambalk, P. De Sutter, E.H. Kostelijk, J.W. Twisk, J. van Rijswijk, E. Van den Abbeel, C.G. Vergouw, Embryo selection using time-lapse analysis (Early Embryo Viability Assessment) in conjunction with standard morphology: a prospective two-center pilot study, Hum. Reprod. 31 (11) (2016) 2450–2457, http://dx.doi.org/10.1093/humrep/dew207.

[23] Y. Miyagi, T. Habara, R. Hirata, N. Hayashi, Predicting a live birth by artificial intelligence incorporating both the blastocyst image and conventional embryo evaluation parameters, Artif. Intell. Med. Imaging 1 (3) (2020) 94–107, http://dx.doi.org/10.35711/aimi.v1.i3.94.

[24] H.A. Güvenir, G. Misirli, S. Dilbaz, O. Ozdegirmenci, B. Demir, B. Dilbaz, Estimating the chance of success in IVF treatment using a ranking algorithm, Med. Biol. Eng. Comput. 53 (2015) 911–920, http://dx.doi.org/10.1007/s11517-015-1299-2.

[25] L. Bori, E. Paya, L. Alegre, T.A. Viloria, J.A. Remohi, V. Naranjo, M. Meseguer, Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential, Fertil. Steril. 114 (6) (2020) 1232–1241, http://dx.doi.org/10.1016/j.fertnstert.2020.08.023.

[26] D.A. Morales, E. Bengoetxea, P. Larranaga, M. García, Y. Franco, M. Fresnada, M. Merino, Bayesian classification for the selection of in-vitro human embryos using morphological and clinical data, Comput. Methods Programs Biomed. 90 (2008) 104–116, http://dx.doi.org/10.1016/j.cmpb.2007.11.018.

[27] A. Uyar, A. Bener, H.N. Ciray, Predictive modeling of implantation outcome in an in vitro fertilization setting: An application of machine learning methods, Med. Decis. Making: An Int. J. Soc. Medical Decis. Making 35 (6) (2015) 714–725, http://dx.doi.org/10.1177/0272989X14535984.

[28] A.A. Septiandri, A. Jamal, P.A. Iffanolida, O. Riayati, B. Wiweko, Human blastocyst classification after in vitro fertilization using deep learning, in: Proceedings of the 7th International Conference on Advance Informatics: Concepts, Theory and Applications, ICAICTA, 2020, http://dx.doi.org/10.1109/ICAICTA49861.2020.9429060.

[29] E. Babayev, Man versus machine in in vitro fertilization—can artificial intelligence replace physicians? Fertil. Steril. 114 (5) (2020) 963, http://dx.doi.org/10.1016/j.fertnstert.2020.07.042.

[30] M. VerMilyea, J. Hall, S.M. Diakiw, A. Johnston, T. Nguyen, D. Perugini, A. Miller, A. Picou, A.P. Murphy, 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, Hum. Reprod. 35 (4) (2020) 770–784, http://dx.doi.org/10.1093/humrep/deaa013.

[31] H. Müller, M.T. Mayrhofer, E.-B. Van Veen, A. Holzinger, The ten commandments of ethical medical AI, Computer 54 (7) (2021) 119–123, http://dx.doi.org/10.1109/MC.2021.3074263.

[32] A. Debón, I. Molina, S. Cabrera, A. Pellicer, Mathematical methodology to obtain and compare different embryo scores, Math. Comput. Modelling 57 (5) (2013) 1380–1394, http://dx.doi.org/10.1016/j.mcm.2012.11.027.

[33] C. Racowsky, L. Ohno-Machado, J. Kim, J. Biggers, Is there an advantage in scoring early embryos on more than one day? Hum. Reprod. 24 (9) (2009) 2104–2113, http://dx.doi.org/10.1093/humrep/dep198.

[34] A.L. Speirs, A. Lopata, M.J. Gronow, G.N. Kellow, W.I.H. Johnston, Analysis of the benefits and risks of multiple embryo transfer, Fertil. Steril. 39 (4) (1983) 468–471, http://dx.doi.org/10.1016/S0015-0282(16)46933-5.

[35] H. Zhou, C.R. Weinberg, Evaluating effects of exposures on embryo viability and uterine receptivity in in vitro fertilization, Stat. Med. 17 (14) (1998) 1601–1612.

[36] S.A. Roberts, C. Stylianou, The non-independence of treatment outcomes from repeat IVF cycles: estimates and consequences, Hum. Reprod. 27 (2) (2012) 436–443, http://dx.doi.org/10.1093/humrep/der420.

[37] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, The MIT Press, 2009.

[38] C. Coughlan, W. Ledger, Q. Wang, F. Liu, A. Demirol, T. Gurgan, R. Cutting, K. Ong, H. Sallam, T.C. Li, Recurrent implantation failure: definition and management, Reproductive BioMedicine Online 28 (1) (2015) 14–38, http://dx.doi.org/10.1016/j.rbmo.2013.08.011.

[39] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1) (1977) 1–38, http://www.jstor.org/stable/2984875.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830, http://jmlr.org/papers/v12/pedregosa11a.html.

[41] O. Chapelle, B. Schökopf, A. Zien, Semi-Supervised Learning, MIT Press, 2006.

[42] J. Hernández-González, I. Inza, J.A. Lozano, Weak supervision and other non-standard classification problems: A taxonomy, Pattern Recognit. Lett. 69 (2016) 49–55, http://dx.doi.org/10.1016/j.patrec.2015.10.008.

[43] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159, http://dx.doi.org/10.1016/S0031-3203(96)00142-2.

[44] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874, http://dx.doi.org/10.1016/j.patrec.2005.10.010.

[45] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning (ICML'06), 2006, pp. 233–240, http://dx.doi.org/10.1145/1143844.1143874.

[46] A.M. Carrington, P.W. Fieguth, H. Qazi, A. Holzinger, H.H. Chen, F. Mayr, D.G. Manuel, A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Med. Inform. Decis. Mak. 20 (2020) 4, http://dx.doi.org/10.1186/s12911-019-1014-6.