Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)

doi:10.3233/FAIA210139

Validation on Real Data of an Extended Embryo-Uterine Probabilistic Graphical Model for Embryo Selection

Adrián TORRES-MARTÍN [a,b], Jerónimo HERNÁNDEZ-GONZÁLEZ [b] and Jesus CERQUIDES [c]
[a] *DEIC, Universitat Autònoma de Barcelona, Bellaterra, Spain*
[b] *Dpt. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain*
[c] *IIIA-CSIC, Campus UAB, Bellaterra, Spain*

Abstract. Embryo selection is a critical step in assisted reproduction (ART): a good selection criteria is expected to increase the probability of inducing pregnancy. In the past, machine learning methods have been used to predict implantation and to rank the most promising embryos. Here, we study the use of a probabilistic graphical model that assumes independence between embryos' individual features and cycles characteristics. It also accounts for a third source of uncertainty attributed to unknown factors. We present an empirical validation and analysis of the behavior of the model within real data. The dataset describes 604 consecutive ART cycles carried out at Hospital Donostia (Spain), where embryo selection was performed following the Spanish Association for Reproduction Biology Studies (ASEBIR) protocol, based on morphological features. The performance of our model is evaluated with different metrics and the predicted probability densities are examined to obtain significant insights about the process. Special attention is given to the relation between the model and the ASEBIR protocol. We validate our model by showing that its predictions show correlation with the ASEBIR score when the score is not provided as a feature. However, once the selection based on this protocol has taken place, our model is unable to separate implanted and failed embryos when only embryo individual features are used. From here, we can conclude that ASEBIR score provides a good summary of morphological features.

Keywords. Assisted Reproductive Technologies, Embryo Selection, Machine Learning, Probabilistic Graphical Models, Learning from Label Proportions

1. Introduction

Assisted reproductive technologies (ARTs) are a set of invasive medical techniques that attempt to induce a pregnancy, used mainly to address infertility. Each trial of a reproduction treatment applying a suitable ART is known as a cycle. When a woman undergoes a cycle, she follows a treatment of ovarian stimulation for several weeks. Then, oocytes are retrieved and fertilized, and the resulting embryos are cultured for several days. Afterwards, the most viable embryos are selected to transfer to the uterus. After transference, the occurrence of embryo implantation determines the process of the cycle. However, for a transfer, current techniques are able to determine the number of embryos that implanted, but unable to identify individually which ones implanted. The probability of pregnancy could be increased by transferring a larger number of embryos [1], but this leads to higher multiple-birth rates, which is considered risky for both mother and the developing fetuses [1, 2]. In fact, in many countries there are legal restrictions limiting the number of embryos 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 their morphological features. Initially, the lack of consensus in this assessment made it impossible to compare results across centres [3]. A unified criteria was created to address this problem: the ASEBIR protocol [4]. This method classifies embryos into a categorical scale (A,B,C,D) using morphological criteria. In recent years, machine learning techniques have been used to assist clinicians in embryo selection and pregnancy prediction [5, 6, 7, 8]. Most of them rely on supervised classification, meaning that only the embryos whose outcome is known (all embryos in the cycles were implanted or none were) are used for training. However, novel methods [7] try to benefit from cycles with partial implantation (not all the transferred embryos were implanted).

The model considered in this paper also deals with cycles with partial implantation. Our model accounts for the factors handled by clinicians in their standard practice as possible determinants of the success of an ART cycle. It assumes independence between embryos and cycles and takes into account a third source of uncertainty corresponding to unknown factors. The main goal of this work is to validate the model using real data, and to study the correlation with ASEBIR score. The paper is organized as follows. Next, we describe the dataset. In Section 2, the model is presented as well as the learning algorithm. In Section 3 the experimental setup is explained, introducing the different probabilistic classifiers and metrics. Then, their results are shown and discussed. Finally, in Section 5 conclusions are drawn and a few open lines for future work are presented.

1.1. Data

The database, originally studied in [7], was collected by the Unit of Assisted Reproduction of the Hospital Donostia (Spain) throughout 18 months (January 2013 - July 2014). It contains 604 cycles of an ART treatment and 3125 associated embryos. Each cycle has a certain number of embryos associated (between 1 and 18), only some of which will actually be transferred (between 1 and 3). Cycles are described by 25 features, mainly related to the female patient, the sperm donor and the stimulation procedure. Embryos are described by 20 features, out of which 13 are morphological features.

Out of the 604 cycles, 192 resulted in a pregnancy with 253 embryos implanted. Of these successful cycles, in 57 of them all the embryos were implanted (108 embryos). In total, the outcome of 947 embryos is known (all embryos implanted in a cycle or none), for 307 we have only the label proportions (in cycle with not all embryos implanted), and for the rest, 1871 embryos, we do not have any information (not transferred embryos).

2. Method

In this work we employ a probabilistic model originally presented in [9] that uses the available information from both cycles and individual embryos, and considers a third source of uncertainty related with unknown factors [10].

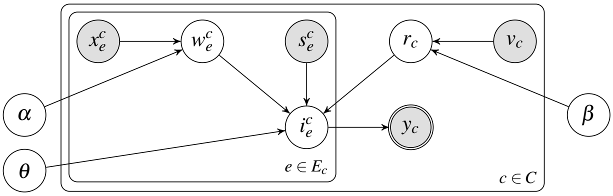

Figure 1. Graphical description of the proposed model. Shadowed nodes represent observed variables. Double line denotes a deterministic variable.

2.1. General probabilistic model

The main assumption of the model is that the probability of an embryo being willing to implant given its own features is independent of the corresponding cycle's features. Similarly, the probability of a cycle being willing to let embryos implant given its own features is independent of the embryos' individual features. Hence embryos and cycles are modeled independently. Moreover, the main novelty is that our model accounts for unknown factors that affect ART success [10] which cannot be explained by the available data. This third source of possible error is included in the model as a Bernoulli distribution with parameter θ_1. The probability of implantation of a high-quality embryo within a cycle willing to let embryos implant is θ_1. If the available information were capable of perfectly predicting the outcome of the process (i.e., no unknown factors), this parameter would be $\theta_1 = 1$. If one of the components (embryo or cycle) is not deemed as good enough to allow implantation then the probability of implantation is directly 0.

Let x_e^c denote the characteristic features of embryo e included in cycle c. Denote by w_e^c a Boolean random variable that represents whether the embryo is willing to implant. This variable w_e^c is modeled by the probability distribution $p(w_e^c \,|\, x_e^c; \alpha)$, where α is the hyperparameter of such distribution. Similarly, let v_c denote the features of cycle c. Denote by r_c a Boolean random variable that represents whether cycle c is willing to let embryos implant, modeled by the probability distribution $p(r_c \,|\, v_c; \beta)$, with hyperparameter β. Both w_e^c and r_c are modeled using probabilistic classifiers.

Let s_e^c denote an observed variable that tells whether embryo e is transferred in cycle c. Denote by i_e^c a Boolean random variable that represents whether embryo e implants in cycle c, modeled by a Bernoulli distribution $i_e^c \sim \text{Bernoulli}(\theta_{w_e^c \cdot r_c \cdot s_e^c})$, given w_e^c, r_c and s_e^c. That is, $\theta_{w_e^c \cdot r_c \cdot s_e^c}$ is only θ_1 when all three variables are positive.

Finally, let y_c denote an observed variable that tells the number of embryos implanted in a cycle. It is just the sum of the i_e^c variables modeling embryo implantation (deterministic), $y_c = \sum_{e \in E_c} i_e^c$, where E_c is the set of embryos associated to cycle c.

Figure 1 shows the complete graphical representation of the model. The shadowed variables are the observed ones (features and final number of implantations per cycle), and θ, α and β are the hyperparameters of the three probability distributions that we are modeling. The other three white nodes w_e^c, i_e^c and r_c represent latent variables, which generally need to be inferred. In some cases the value of y_c is enough to deduce the value of these variables. For example, if $y_c > 0$ then we know that this cycle is willing to let embryos implant ($r_c = 1$). However, if $y_c = 0$ we do not know which was the actual cause of failure: the embryo, the cycle or an unknown factor. The complete joint probability is

$$p(\mathbf{x},\mathbf{w},\mathbf{v},\mathbf{r},\mathbf{s},\mathbf{i},\mathbf{y};\alpha,\beta,\theta) = p(\mathbf{w}|\mathbf{x};\alpha)p(\mathbf{x})p(\mathbf{r}|\mathbf{v};\beta)p(\mathbf{v})p(\mathbf{s})p(\mathbf{y}|\mathbf{i})p(\mathbf{i}|\mathbf{w},\mathbf{r},\mathbf{s};\theta) \quad (1)$$

The goal of the learning algorithm is to estimate the set of hyperparameters parameters θ, α and β that maximize the conditional probability:

$$p(\mathbf{y}|\mathbf{x},\mathbf{v},\mathbf{s};\alpha,\beta,\theta) = \sum_{\mathbf{r}} p(\mathbf{r}|\mathbf{v};\beta) \sum_{\breve{i}\in\mathbb{I}_{\mathbf{s},\mathbf{y}}} \sum_{\mathbf{w}} p(\breve{\mathbf{i}}|\mathbf{w},\mathbf{r},\mathbf{s};\theta)p(\mathbf{w}|\mathbf{x};\alpha) \tag{2}$$

where $\mathbb{I}_{\mathbf{s},\mathbf{y}}$ is the set of vectors i compatible with the selections $\{s_e^c\}$ and the known outcomes $\{y_c\}$. E.g., in a cycle with 4 embryos, where only the first and third are selected and only one of them was implanted, the possible vectors are $[1,0,0,0]$ and $[0,0,1,0]$.

2.2. EM Algorithm

In the presented model there are latent variables (w_e^c, i_e^c and r_c) whose value is generally unknown. We use an Expectation-Maximization (EM) algorithm [11] to learn in this scenario, combining the completion (expectation) of these latent variables with the estimation of the hyperparameters θ, α and β maximizing the log-likelihood.

For each cycle c we consider a pair of weights $q(r_c = r)$ associated to the two possible values of r_c, $r \in \{0,1\}$. These weights are computed as the likelihood of obtaining $r_c = r$ taking into account the whole model, not just the features of the cycle:

$$q(r_c = r) \propto \Big(\sum_{\mathbf{i}^c \in \mathbb{I}_{\mathbf{s}_:^c, y_c}} \prod_e \sum_{w_e^c} p(i_e^c|w_e^c, r_c = r, s_e^c;\theta)p(w_e^c|x_e^c;\alpha)\Big) p(r_c = r|v_c;\beta). \tag{3}$$

Similarly, for each embryo e in the cycle we compute the weights corresponding to the two values of w_e^c, $w \in \{0,1\}$:

$$q(w_e^c = w) \propto \sum_{r_c} \left(\sum_{\mathbf{i}^c \in \mathbb{I}_{\mathbf{s}_:^c, y_c}} p(i_e^c|w, r_c, s_e^c;\theta)p(w|x_e^c;\alpha) \cdot \right. \\ \left. \prod_{e' \neq e} \sum_{w_{e'}^c} p(i_{e'}^c|w_{e'}^c, r_c, s_{e'}^c;\theta)p(w_{e'}^c|x_{e'}^c;\alpha) \right) p(r_c|v_c;\beta) \tag{4}$$

Finally, the weights associated to each possible combination ($\mathbf{i} \in \mathbb{I}_{\mathbf{s}_:^c, y_c}$) for $\mathbf{i}^c$ are:

$$q(\mathbf{i}^c = \mathbf{i}) \propto \sum_{r_c} \Big(\prod_e \sum_{w_e^c} p(i_e|w_e^c, r_c, s_e^c;\theta)p(w_e^c|x_e^c;\alpha)\Big) p(r_c|v_c;\beta) \tag{5}$$

Our algorithm starts with the initialization, where the weights are randomly assigned and normalized (to sum up to 1). If the real value of the variable is known, these values are fixed (e.g., if $y_c > 0$ then $q(r_c = 1) = 1$ and $q(r_c = 0) = 0$). Then, it repeats iteratively:

Expectation: The unfixed weights are updated with Equations 3, 4 and 5, using the current fit of the model $(\hat{\alpha}, \hat{\beta}, \hat{\theta}_1)$.

Maximization: Hyperparameters $(\alpha, \beta, \theta_1)$ are re-estimated. For α and β, we retrain the probabilistic classifiers with the new weights obtained from the previous E-step. For θ_1, the maximum likelihood estimator is:

$$\hat{\theta}_1 = \frac{\sum_c \sum_{\mathbf{i}^{c'} \in \mathbb{I}_{\mathbf{s}_{:}^{c}, y_c}} \sum_e q(\mathbf{i}^{c'}) q(r_c = 1) q(w_e^c = 1) i_{e'}^c}{\sum_c \sum_{\mathbf{i}^{c'} \in \mathbb{I}_{\mathbf{s}_{:}^{c}, y_c}} \sum_e q(\mathbf{i}^{c'}) q(r_c = 1) q(w_e^c = 1)} \tag{6}$$

The method iterates until the stopping condition is met (maximum number of iterations or convergence of weights). It is run multiple (10) times with different initialization to mitigate the local-maximum problem of EM algorithms.

3. Experimental setup

The main goal of this project is to analyze the performance of the general probabilistic model proposed in Section 2.1 and to compare the performance with different probabilistic classifiers. Moreover, we study the effect in our model of the ASEBIR quality score [4], and whether both our model and this score agree on the embryo selection.

Our model uses probabilistic classifiers to predict the probability that an embryo is willing to implant, $p(\mathbf{w}|\mathbf{x};\alpha)$, and that a cycle is willing to let embryos implant, $p(\mathbf{r}|\mathbf{v};\beta)$. Different classifiers may perform differently depending on the context. In order to make a fair comparison between the various models we use three different probabilistic classifiers: (i) Extra-trees classifier (ETREES) (ii) Gradient Boosting (GBOOST) and (iii) Logistic Regression (LR).

Because of the weakly supervised nature of the problem [12], the evaluation of the models is not trivial and needs to be properly addressed in order to make a fair comparison. For instance, a large fraction of embryos in the dataset were not actually transferred; hence they are not labeled as implanted or not. We use them in our EM learning algorithm but they cannot be used to assess model performance. Moreover, a part of the transferred embryos have no label: when only some of the embryos in their cycle were implanted. However, for these, we do know the proportion of the embryos that were implanted. This information should be used to take full advantage of the data.

Also the interpretation of the predictions needs proper consideration. For instance, the full model gives the probability of implantation of an embryo in a certain cycle assuming independence between embryo and cycle. In fact, we can compute the probability of both embryos and cycles of being appropriate for ART directly with the respective probability classifier. This means that, if we only want to rank a set of embryos according to their *quality*, we could use just the embryo classifier trained within the whole model.

To test the performance of the model and obtain relevant metrics and probability densities, we use 5-fold cross validation. The resulting measures are averaged to obtain a final evaluation metric. Most of the metrics used here need the probability of implantation of an embryo in a cycle, which is given by:

$$p(i_e^c = 1|x_e^c, s_e^c, v_c; \alpha, \beta, \theta) = p(i_e^c = 1|w_e^c = 1, s_e^c, r_c = 1; \theta) p(w_e^c = 1|x_e^c; \alpha) p(r_c = 1|v_c; \beta). \tag{7}$$

where $p(i_e^c = 1|w_e^c = 1, s_e^c, r_c = 1; \theta) = \theta_1 \cdot s_e^c$. Remember that if $s_e^c = 0$, $p(i_e^c = 1|w_e^c, s_e^c = 0, r_c; \theta) = 0$. That is why the evaluation is only performed with embryos which were transferred ($s_e^c = 1$). The other two terms in Eq. 7 are given by the probabilistic classifiers.

Performance is assessed in terms of different metrics. To test the ability to predict embryo implantation, we use only the embryos whose fate is known (i.e., those belonging to completely implanted cycles or failed cycles) and measure the AUC-ROC [13].

Table 1. Metrics and control measures obtained using 5-fold cross validation

Classifier	AUC-ROC	LP-loss	loglikelihood	AUC-ROC	LP-loss	loglikelihood
ETREES	0.64 ± 0.07	**0.54 ± 0.05**	1.45 ± 1.59	0.64 ± 0.05	**0.54 ± 0.05**	1.27 ± 1.57
GBOOST	**0.71 ± 0.04**	0.72 ± 0.03	**0.45 ± 0.05**	**0.73 ± 0.07**	0.73 ± 0.07	**0.43 ± 0.06**
LR	0.63 ± 0.08	0.60 ± 0.05	0.51 ± 0.10	0.62 ± 0.08	0.64 ± 0.07	0.52 ± 0.10
	Full model			Full Model, hidden quality		

To account also for the partially implanted cycles, we use the label proportion loss (LP-loss) and the negative log-likelihood. LP-loss measures how close the real and predicted label proportions are. For each cycle, the difference between the number of embryos predicted as implanted and the actual number of implanted embryos is taken in absolute value. The LP loss is the mean value of these differences. Similarly, we might want to consider how confident is the model in predicting each of these labels. For that matter, we use the negative log-likelihood. As most of the embryos do not have a true label to compare with, we compute this measure cycle by cycle, calculating the likelihood of the real number of implanted embryos within the learnt model. Let N_c be the number of transferred embryos in cycle c, and y_c the number of implanted ones. The negative log-likelihood is

$$\mathscr{L}(\mathbf{Y};\alpha,\beta,\theta) = -\frac{1}{B}\sum_{c}\sum_{j=0}^{N_c} \mathbb{1}[y_c = j]\log p(y_c), \tag{8}$$

where B is the total number of cycles and $p(y_c)$, the probability of cycle c having y_c implanted embryos, is,

$$p(y_c) = \sum_{i^c \in \mathbb{I}_{y_c}} \prod_{e} [i^c_e p(i^c_e = 1) + (1 - i^c_e) p(i^c_c = 0)] \tag{9}$$

where $p(i_c)$ is given by Eq. 7 and $\mathbb{I}_{y_c}$ consists of the possible joint assignment of value (vector) to all the $\{i^c_e\}_{e \in E_c}$, as explained in the context of Eq. 2.

4. Results and discussion

A relevant point is whether our model agrees with the ASEBIR score. In our dataset, we have this score as a feature, as well as all the factors used to compute it. To study the agreement, we trained the model in two different ways: with and without this quality score included as a feature of embryos. In Table 1 we show the metrics obtained for each probabilistic classifier and for both models (with and without ASEBIR score feature). Observe that there are no significant differences between the two different models. The model seems not to be directly using the feature as a discriminant for implantation. It must be gathering that information from the other features in the dataset which are, in fact, the ones used in their protocol [4].

In terms of performance, GBOOST seems to be the best one according to AUC-ROC and negative log-likelihood. ETREES and LR classifiers are both similar regarding AUC-ROC but their log-likelihood values are rather different. A critical difference between these two measures is that they use a different set of embryos for evaluation. AUC-ROC

Table 2. Estimated parameter θ_1 for the three different classifiers.

Model	Classifier	θ_1	Model	Classifier	θ_1
Full Model	ETREES	0.60 ± 0.04	Full Model (Hidden quality)	ETREES	0.58 ± 0.04
	GBOOST	0.49 ± 0.00		GBOOST	0.49 ± 0.01
	LR	0.52 ± 0.01		LR	0.51 ± 0.00

is calculated using only embryos whose outcome is known, whereas log-likelihood uses all transferred embryos, evaluating cycle by cycle the proportion of implanted embryos. Thus, ETREES performs relatively well in a pure classification task (implantation or not), but it fails on estimating the probability of more uncertain cases.

Table 2 shows the mean estimation of the parameter θ_1 obtained with each classifier and model, over the different CV folds. The standard deviation is quite low for all the classifiers, implying a consistent estimation. This parameter is the probability that a good embryo will actually get implanted in a good cycle. It represents the third source of failure for implantation of our model, and accounts for all unknown factors.

For the GBOOST and LR classifiers, the mean value of θ_1 is close to 0.5. This means that these models, even when the classifiers consider that both embryo and cycle are promising, expect that only half of these pairs will succeed. The ETREES classifier estimates a noticeably higher $\theta_1 = 0.58$. This might suggest that this model has a higher confidence on the judgement of its embryo and cycle classifiers. Unfortunately, this confidence does not translate into better results (see Table 1).

To fully grasp the behaviour of the models, we also analyze the different predicted probability densities output by them. Figure 2 displays the densities, separated for successful and failed cycles, of (i) whether the embryo is willing to implant, (ii) whether the cycle is willing to accept embryos, and (iii) whether the ART treatment is leading to a pregnancy (whole model). The ideal classifier would separate clearly the curves of each class for the third case (right column). The results of all classifiers are quite similar: Although intersection between both densities is considerable, the mode of the density for successful treatments (pregnancy) is clearly to the right regarding that of the failed treatments. This means that, on average, *the models predict the actual implanted embryos as more likely to implant than the failed ones*.

In the first column of Figure 2, the probability of deeming an embryo as willing to implant is practically the same for successful and failed treatments. At a first sight, one could think that embryos are not relevant to predict a pregnancy. Nevertheless, it is noteworthy that the embryos employed in this study are only the ones that were transferred. And, transferred embryos are usually the best embryos as selected by the embryologists, that is, all the embryos that we observed were considered as good-quality ones by the specialists. Instead, most of the predictive power seems to come from the cycle. In the middle column of Figure 2, it can be observed that the classifier gives a higher probability of being a cycle willing to be implanted to those treatments that induced a pregnancy. All this could mean that *the protocol followed by the embryologists performs well in selecting the best embryos based on the morphological features*. Our model is not able to further discriminate the embryos based on this data (the same they used) alone.

Figure 3 displays, in a similar way, different probability densities separating on the ASEBIR score of the involved embryo. For this experiment, we hide the ASEBIR score feature from the model. Under the independence hypothesis, the quality of an embryo should not affect the probability that a cycle is in good conditions and, for the most part

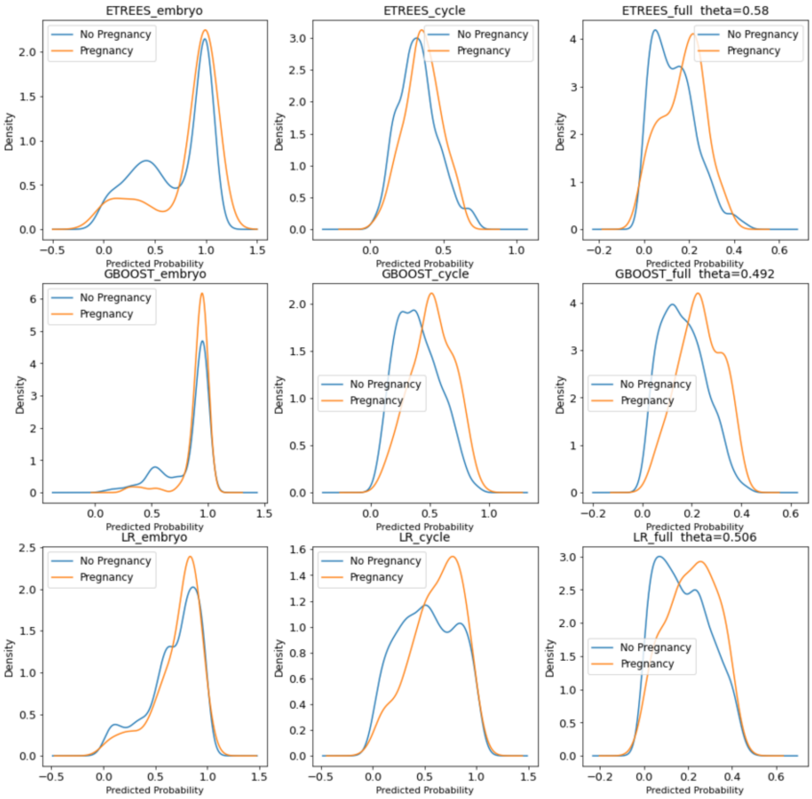

Figure 2. Density of the predicted probabilities for an embryo to be willing to implant, for a cycle to be willing to let embryos implant and for the pair embryo-cycle to actually induce pregnancy. The figure shows the different probability densities depending on the true outcome of each embryo-cycle pair (induce pregnancy or not). Each row corresponds to a different probabilistic classifier (ETREES, GBOOST and LR).

of it, we observe that the embryo information has not leaked into the cycle classifier. However, with ETREES there is a slight disparity in favor of treatments using embryos of good quality. Embryo quality has the highest impact on the probability of considering an embryo as willing to implant. All classifiers separate quite well the best (A) and worst (D) quality embryos. ETREES and GBOOST seem not to differentiate embryos of medium quality (B and C) completely, while LR does separate them slightly. For all classifiers the embryo quality does translate well into the final prediction of implantation. Note that this does not validate the model regarding implantation, but it implies that *the model mostly agrees with the ASEBIR score in the selection of the most promising embryos based on this set of features*.

5. Conclusions

In this paper, we address embryo selection for ARTs using a probabilistic model that assumes independence between embryos and cycles. Using morphological data for each individual embryo and characteristics about the cycle, the model is able to predict im-

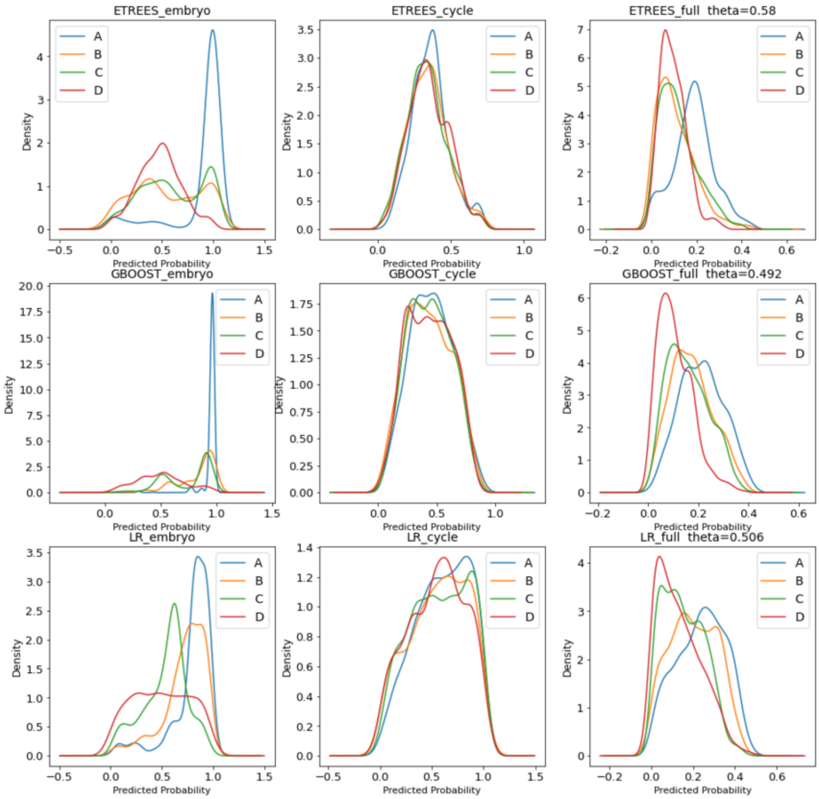

Figure 3. Density of the predicted probabilities for an embryo to be willing to implant, for a cycle to be willing to let embryos implant and for the pair embryo-cycle to actually implant. The figure shows the different probability densities depending on the ASEBIR quality score given to the embryo (A, B, C or D). Each row corresponds to a different probabilistic classifier (ETREES, GBOOST and LR).

plantation. The performance of the model is tested using different classifiers which evaluate the goodness of the embryos and cycles. The Gradient Boosting classifier showed the best results both in terms of AUC-ROC and negative log-likelihood.

The probability densities obtained from the predictions provided helpful insights to understand the behaviour of the model. We studied the effect of the ASEBIR embryo quality score within our model. We have not observed differences between models learnt with and without the ASEBIR score directly as a feature. The probability densities grouped by this quality feature show a clear separation between groups (especially between the best and worst grades), using both models. We have observed that, once embryologists made their selection, the model does not provide more information about individual embryos. This might indicate that the protocol followed by the embryologists is already extracting most of the value out of the morphological data.

There are different research lines open after this exploration of the behaviour of the model in relation to the ASEBIR protocol. We plan to enlarge our experimental setup to obtain a deeper understanding of the intricacies of the model. Another direction would be to conceive new, maybe simpler, models to test the assumptions of our current model (independence between embryos and cycles, awareness of a third source of error, etc.).

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. We would like to thank Olga Valls for providing the source code of the model, and the physicians from Hospital Donostia (Spain) who collected the data and granted us permission to use it.

References

[1] L Engmann, N Maconochie, S Tan, and J Bekir. Trends in the incidence of births and multiple births and the factors that determine the probability of multiple birth after IVF treatment. *Human Reproduction*, 16:2598–605, 12 2001.

[2] ESHRE Campus Course Report. Prevention of twin pregnancies after IVF/ICSI by single embryo transfer. *Human Reproduction*, 16(4):790–800, 04 2001.

[3] I Cuevas-Siz, M C Pons, M Vargas, A Mendive, N Enedáguila, M Solanes, B Carrasco, J López, A Bonet, and M Acosta. The Embryology Interest Group: updating ASEBIR's morphological scoring system for early embryos, morulae and blastocysts. *Medicina Reproductiva y Embriologa Clnica*, 5, 02 2018.

[4] M. Ardoy and G. Calderon. Clinical embryology papers: Asebir criteria for the morphological evaluation of human oocytes, early embryos and blastocysts. *Asociacin para el Estudio de la Biologa de la Reproduccin (ASEBIR)*, 2008.

[5] G Corani, M C Magli, A Giusti, L Gianaroli, and L M Gambardella. A bayesian network model for predicting pregnancy after in vitro fertilization. *Computers in biology and medicine*, 43:1783–92, 11 2013.

[6] F. Guérif, A. le Gouge, B. Giraudeau, J. Poindron, R. Bidault, O. Gasnier, and D. Royère. Limited value of morphological assessment at days 1 and 2 to predict blastocyst development potential: a prospective study based on 4042 embryos. *Human reproduction*, 22 7:1973–81, 2007.

[7] J Hernández-González, I Inza, L Crisol-Ortíz, M A Guembe, M J Iñarra, and J A Lozano. Fitting the data from embryo implantation prediction: Learning from label proportions. *Statistical Methods in Medical Research*, 27:1056 – 1066, 2018.

[8] M Kragh, J Rimestad, J Berntsen, and H Karstoft. Automatic grading of human blastocysts from time-lapse imaging. *Comput. Biol. Med.*, 115:103494, 2019.

[9] O Valls Murcia. A comprehensive probabilistic model for the embryo selection problem. Master's thesis, Technical University of Catalonia, 2021. URL `http://hdl.handle.net/2117/340945`.

[10] C Coughlan, W Ledger, Q Wang, F Liu, A Demirol, T Gurgan, R Cutting, K Ong, H Sallam, and T C Li. Recurrent implantation failure: definition and management. *Reproductive BioMedicine Online*, 28(1):14–38, 2015.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[12] J Hernndez-Gonzlez, I Inza, and J A Lozano. Weak supervision and other non-standard classification problems: A taxonomy. *Pattern Recogn. Lett.*, 69:49–55, 2016.

[13] T Fawcett. Introduction to ROC analysis. *Pattern Recogn. Lett.*, 27:861–874, 2006.