Fitting the data from embryo implantation prediction: learning from label proportions

Jerónimo Hernández-González*[1], Iñaki Inza[1], Lorena Crisol-Ortíz[2], María A. Guembe[2], María J. Iñarra[2], and Jose A. Lozano[1,3]

[1]Intelligent Systems Group, University of the Basque Country UPV/EHU, Spain
[2]Unit of Assisted Reproduction, Osakidetza - Basque Public Health Service, Spain
[3]Basque Center for Applied Mathematics BCAM, Spain

Abstract

Machine learning techniques have been previously used to assist clinicians to select embryos for human assisted reproduction. This work aims to show how an appropriate modeling of the problem can contribute to improve machine learning techniques for embryo selection. In this study, a dataset of 330 consecutive cycles (and associated embryos) carried out by the Unit of Assisted Reproduction of the Hospital Donostia (Spain) throughout 18 months has been analyzed. The problem of the embryo selection has been modeled by a novel weakly supervised paradigm, learning from label proportions, which considers all the available data, including embryos whose fate cannot be certainly established. Furthermore, all the collected features, describing cycles and embryos, have been considered in a multi-variate data analysis. Our integral solution has been successfully tested. Experimental results show that the proposed technique consistently outperforms an equivalent approach based on standard supervised classification. Embryos in this study were selected for transference according to the criteria of the Spanish Association for Reproduction Biology Studies. Obtained classification models outperform this criteria, specifically reordering medium-quality embryos.

Assisted reproductive technologies, Embryo selection, Machine learning, Learning from label proportions, Bayesian network models

*Faculty of Informatics, Lab. 309, P. Manuel Lardizabal 1, 20018 Donostia, Spain. Email: jeronimo.hernandez@ehu.eus; Tel.: (+34) 943 018 070

1 Introduction

Assisted reproductive technologies (ARTs) are a set of invasive medical techniques that attempt to induce a pregnancy. Each trial of a reproduction treatment applying a suitable ART is known as a cycle. When a woman undergoes an ART cycle, she follows a treatment of ovarian stimulation for several weeks in order to induce the development of multiple follicles with a large number of oocytes. By means of a punction, oocytes are retrieved using an ultrasound-guided transvaginal follicle aspiration. The mature oocytes are subsequently fertilized and the resulting embryos cultured for several days. In a critical decision, clinicians usually have to select the embryos to transfer as the clinical procedure can produce excess embryos. Although the number of transferred embryos is positively correlated with the probability of pregnancy [1, 2], multi-transference may give room to a *multiple* pregnancy, which is widely considered risky for both the woman and the developing fetus(es) [1, 2, 3, 4, 5]. In order to reduce the occurrence of multiple pregnancy, legal restrictions limiting the maximum number of transferred embryos have been established (e.g., Spanish law limits it to 3). Not only the number but also the individual embryos to transfer have to be carefully selected as the transference of poor-quality embryos is a major contribution to ART failure [6, 7]. After transference, the occurrence of embryo implantation —a natural process that cannot be monitored by the specialist— determines the success of an ART cycle: Implantation of at least one of the transferred embryos leads the cycle to a pregnancy.

For decades, there has been a persisting discussion on the features that determine the success of a cycle. In their exhaustive reviews, Achache and Revel [6] and Ebner et al. [7] collected and discussed an extensive set of features that have been considered for assessing the quality of both cycles and oocytes/embryos. Many research works have made use of data analysis techniques to determine the contribution of specific features. Similarly, an unbounded number of embryo scores and selection criteria have been presented [7, 8, 9, 10, 11, 12, 13]. More recently, taking advantage of the development of computational techniques, different machine learning (ML) paradigms have been applied to the analysis of the ART problem [14, 15, 16, 17, 18, 19, 20].

In this paper, the ML paradigm of supervised classification is considered to deal with the *implantation* prediction problem, i.e., to classify embryos according to their probability of resulting implanted or not if they were transferred. In the most popular ART application of ML techniques [14, 15, 16, 18, 19], the supervised classification framework, a classification model that reproduces the inherent categorizing behavior of a problem of interest is built/learnt from a set of previous examples. Each example describes a real case of the problem and has been annotated with its real category (class label). The classifier predicts (accurately) the category of new uncategorized examples. However, in our problem the previous evidence is composed of transferred embryos which cannot always be certainly categorized: Current medical techniques only allow clinicians to know the *number* of implanted embryos, not their *identity*. That is, the individual *fate* of the embryos for training is usually unknown. It can

be known only in those cases where none or all the transferred embryos were implanted. In practice, many studies in the related literature discarded the embryos of unknown fate [10, 12, 13, 15, 19, 21] for model learning. However, as the conclusions drawn from the results of any data analysis are affected by the sample size, ML researchers have focused their effort on incorporating any available example to the analysis. This has led to the proposal of novel ML techniques which use all the available —even incomplete— information to train the classification models. This is the case of the recently proposed learning from label proportions (LLP) [22] paradigm, where the training dataset is divided in groups (*bags*) of unlabeled examples and, for each group, the number of examples of each category is provided. Previous studies show that classification models can be efficiently learnt from this kind of partially labeled data [22, 23]. The embryo implantation prediction problem naturally fits the LLP paradigm: each group of embryos transferred in the same ART cycle forms a bag and, for each bag, the number of implanted embryos is known.

Another novelty of this work also involves the use of any available information: All the collected features have been considered. Bayesian network classifiers [24], a type of model which can be calibrated to balance the contribution of each predictive feature, have been used. However, when using a larger set of features, irrelevant/redundant features can be introduced, which is usually harmful to the performance of the classifier [25]. Thus, feature subset selection (FSS) techniques [26, 27] have been used to automatically identify the relevant predictive features and discard those that are uninformative and/or redundant. With our strategy, physicians do not need to manually select the features: The method inputs all the collected features and automatically establishes their contribution.

In this paper, the implantation prediction problem is modeled by means of the learning from label proportions problem. The proposed methodology is described in the following section. Next, a large set of experiments is presented and their results are discussed. Finally, conclusions are drawn and future research lines are noted.

2 Material and methods

2.1 Data

The database has been collected by the Unit of Assisted Reproduction of the Hospital Donostia (Spain) throughout 18 months (January 2013 - July 2014). The population consisted of 330 consecutive patients participating in the IVF-ICSI program and a total of 696 embryos. As detailed in Table 1, 217 cycles failed to induce a pregnancy (all the 447 embryos were not implanted) and in 39 cycles all the transferred embryos resulted implanted (i.e., 72 implanted embryos). In each of the remaining cycles (74), where only a subset of the transferred embryos became implanted, the fate of each individual embryo cannot be claimed (in total, 117 embryos). The database is composed of two spreadsheets,

one for *cycles* and another for *embryos*, related by a one-to-n relationship: one cycle, the n embryos transferred in that procedure. Each cycle is described by 26 features, including characteristics of the patient, stimulation treatment and statistics of the associated embryos. Additionally, 14 features are used to describe embryos: oocyte/embryonic morphological characteristics and quality grades. A complete description of the database can be found in the webpage associated with this paper[1].

2.2 Protocol

The IVF management mostly consisted of GnRH antagonist protocol. Briefly, the suppression of pituitary FSH and LH secretion was performed with 0.25 mg cetrorelix (Cetrotide, Asta Medica) administered daily when two or more follicles reached 13–14 mm in diameter. Occasionally, down-regulation with a GnRH analogue, triptorelin acetate (Synarel, Lab. Seid) on a long protocol was performed. Ovarian stimulation was performed with recombinant FSH (Gonal F, Merck Serono), highly purified urinary FSH (Angelini) or highly purified urinary menopausal gonadotropins (Menopur, Ferring) depending on each patient's characteristics. The doses of hMG and FSH were adjusted according to the ovarian response. Ovulation was triggered with 250 mg Ovitrelle (Merck Serono) and transvaginal ultrasound-guided oocyte retrieval was scheduled 36 hours after hCG injection. Oocytes were inseminated 4 hours after retrieval, ordinarily using conventional IVF. ICSI was performed in cases with less than 1.5 million motile sperm recovered after capacitation, low fertilization rate ($< 30\%$) in previous IVF cycles, and/or previous intrauterine insemination failures. Embryo transfer was performed on day 2. The embryo selection criteria of the Association for Reproduction Biology Studies [28] (Spanish acronym ASEBIR), extensively used by Spanish clinicians, was followed. The luteal phase was vaginally supplemented with 200 mg micronized progesterone (Utrogestan, Lab. Seid, or Progeffik, Effik) every 12 hours. Pregnancy test was carried out 14 days after embryo transference.

2.3 Paradigm: Learning from label proportions

In supervised classification, a problem is described by a set of n predictive features $(X_1, \ldots, X_n)$ and a special feature, the class variable C. Each of them

[1] http://www.sc.ehu.es/ccwbayes/members/jeronimo/llp_embryo/

		Implanted			
		0	1	2	3
Transfer	1	32	8	-	-
	2	140	45	29	-
	3	45	20	9	2

Table 1: Summary table with the number of cycles according to their number of transferred/implanted embryos.

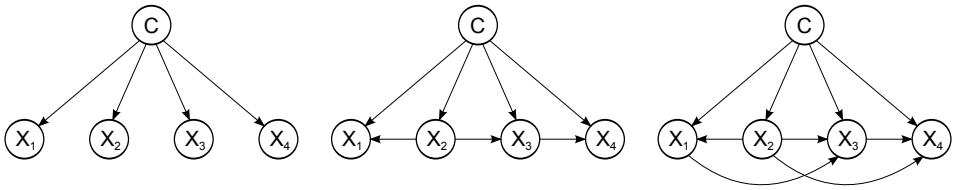


Figure 1: Examples of the three types of BNC structures used in this study. From left to right, NB, TAN and 2DB models are allowed to capture, respectively, a maximum number of 0, 1 or 2 conditional dependencies between predictive variables.

has a set of possible values. Specifically, the term "class label" refers to each possible value of the class variable and $\mathcal{C}$ represents the set that groups all the class labels. Thus, a problem example $(\boldsymbol{x}, c)$ is a $(n+1)$-tuple that assigns a value to each feature. The objective is to learn from a set of previous examples a classification model which infers the class label (category) of new unclassified examples. The training set is composed of m *fully* labeled examples $D = \{(\boldsymbol{x}_1, c_1), (\boldsymbol{x}_2, c_2), \ldots, (\boldsymbol{x}_m, c_m)\}$, which are supposed to be i.i.d. sampled from some underlying probability distribution.

The implantation prediction problem is modeled by means of the learning from label proportions (LLP) paradigm. Although it has the same objective and problem description as the standard supervised classification, it does not provide a completely labeled training dataset. The m training examples are individually unlabeled and the dataset is divided in b *bags* $D = \boldsymbol{B}_1 \cup \boldsymbol{B}_2 \cup \cdots \cup \boldsymbol{B}_b$, where $\boldsymbol{B}_i \cap \boldsymbol{B}_j = \emptyset, \forall i \neq j$. A bag $\boldsymbol{B}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{im_i}\}$ groups m_i instances $(\sum_{i=1}^{b} m_i = m)$ and provides limited class information: the m_{ic} values or *counts* $(\sum_{c \in \mathcal{C}} m_{ic} = m_i)$ which indicate the number of instances in $\boldsymbol{B}_i$ that belong to class label c.

2.4 Classification models

In this analysis, Bayesian network models are used as probabilistic classifiers (BNC) [24]. A Bayesian network, represented by a pair $(G, \boldsymbol{\theta})$, is a probabilistic graphical model that encodes the conditional (in)dependencies between a set of random variables (features) using a directed acyclic graph. The graph structure, $G = (\mathcal{V}, \mathcal{R})$, codifies the arcs $\mathcal{R}$ —conditional (in)dependencies— between the random variables $\mathcal{V} = (X_1, \ldots, X_n, C)$, and $\boldsymbol{\theta}$ is the set of parameters of the conditional probability functions of each random variable given its parents in the graph.

The outstanding interpretability of Bayesian network models has motivated our choice: influences and dependencies among random variables can be induced from the explicit probability relationships. In the context of the ARTs, several authors have already used BNCs as probabilistic classifiers [14, 17]. Specifically, three kinds of Bayesian network classifiers have been considered: naive

Bayes (NB) [29], tree augmented naive Bayes (TAN) [30] and K-dependence Bayesian network (KDB) [31]. Based on the assumption of conditional independence between the predictive variables given the class variable, the naive Bayes presents the simplest network structure (see Figure 1). TAN and KDB are the next step forward in terms of network structure complexity and allow models to capture some conditional dependencies between predictive variables. The general classification rule of these BNCs is defined as:

$$\underset{c}{\operatorname{argmax}}\, p(C=c)\prod_{v=1}^{n} p(X_v = x_v|\mathbf{PA}_v = \boldsymbol{pa}_v, C=c)$$

where $\boldsymbol{pa}_v$ is the vector of values assigned in the example $\boldsymbol{x}$ to the predictive variables $\mathbf{PA}_v$ which are parents of X_v in G. In the case of NB classifiers, the set $\mathbf{PA}_v$ is always empty, $\mathbf{PA}_v = \emptyset$. In TAN structures, it usually has size $|\mathbf{PA}_v| = 1$, with the only exception of the root variable of the tree, which does not have predictive variables as parents, $\mathbf{PA}_r = \emptyset$. K is a constraint in the number of predictive variables used as parents in KDB structures, $|\mathbf{PA}_v| \leq K$.

From a set of labeled examples, both the model parameters and the graph of conditional (in)dependencies of a BNC can be estimated. Maximum likelihood estimates of the model parameters can be obtained by means of frequency counts [32]. Regarding the graph structure, Friedman et al. [30] and Sahami [31] proposed methods for learning TAN and KDB structures, respectively. As the naive Bayes structure is fixed, no structural learning is required.

To learn from a partially labeled set such as that collected for the implantation prediction problem, alternative procedures are required. In this paper, a method [22] for learning BNCs from this kind of data based on the Expectation-Maximization (EM) [33] strategy has been used. It consists of an iterative procedure that combines the completion of the partially labeled dataset and model improvement. At each iteration and for each bag, the most probable probabilistic labeling is calculated and, using it to complete the data, the model is updated.

2.5 Experimental Setting

A complete experimental setting has been designed. Two training datasets have been derived from the gathered database. First, a dataset where the cases (embryos) are just described by embryonic features. The second dataset combines embryonic and cycle features to describe the embryos. In both cases, the class "implantation" is a binary feature (positive and negative are the two possible values or class labels) which represents the fate of the embryo (implanted or not implanted, respectively). That is, an embryo is considered as "positive" if it resulted implanted after transference and "negative" otherwise. All the experiments are repeated for both datasets. The exposed three types of BNCs (Fig. 1) have been learnt for each experimental configuration. All the continuous features have been discretized using equal-frequency with 3 intervals. With respect to FSS, a multivariate and a univariate strategy have been used. The former

applies the popular correlation-based feature subset selection [34] (with both backward and forward search strategies) to obtain a subset of non-redundant predictive features highly correlated with the class. It has been carried out using a dataset completed according to the label proportions of the bags. The latter carries out chi-square statistical tests between the class and each predictive feature, and uses the resulting p-values to build an order of relevant predictive features. Different experiments have been carried out using the subset of the s most relevant features ($s \in \{n, \ldots, max(n_p, 2)\}$, where n_p is the number of predictive features with a p-value ≤ 0.05). Taking advantage of embryos in *full* bags (those that represent cycles where all or no embryo became implanted, i.e., they are actually labeled), the ranking of relevant features has been calculated in a pre-process stage. These labeled embryos have been also used for evaluation. A leave-one-*full-bag*-out procedure has been used, which takes a full bag at each iteration as the validation set and the rest of bags for model training.

In order to study the performance gain derived from the use of the LLP paradigm and its fitted modeling of the embryo implantation prediction problem, all the experiments have been replicated to learn the same type of classifiers with the standard supervised classification paradigm. Note that this standard paradigm is the methodology considered by previous machine learning approaches to the embryo selection problem. Following the general procedure to prepare the dataset for standard learning techniques [10, 12, 13, 15, 19, 21], all the embryos in non-full bags (embryos of unknown fate) are removed. Thus, a fully labeled dataset is obtained and the aforementioned state-of-the-art learning techniques are used to learn NB [29], TAN [30] and 2DB [31] classifiers.

Additionally, the best performing configuration of each method and classifier has been evaluated with an auxiliary dataset in order to show the generalization ability of the learnt models in reserved data. The dataset consists of 134 cycles carried out by our team of physicians from August 2014 to June 2015. Only cycles where all the transferred embryos had the same fate (full bags) —all of them resulted implanted or failed to implant— have been considered. It involves 253 embryos, from which 45 resulted implanted and 208 failed to implant. Gathered in the same conditions, the reserved dataset is consecutive and exclusive with respect to the dataset used for the learning process.

The objective of these experiments is to show why it is worth using a fitted modeling that takes full advantage of it. Alternatives to enhance the performance of the learnt classifiers (e.g., probabilistic classifiers with the decision boundary not in 0.5 or loss functions that penalize asymmetrically false positives and negatives) have not been considered. To put them in production supporting the decision making process of the physicians in their daily practice, an in-depth learning methodology to maximize a specific quality metric —subject to the specialist's preferences— should be considered. However, this is beyond the scope of the present work.

3 Results

Table 2 shows the results of the experiments where the three types of BNCs considered in this study (NB, TAN and 2DB; Fig. 1) were learnt from both exposed datasets (using as predictive features only embryonic features or a combination of embryonic and cycle features). The results of the BNCs learnt in the LLP paradigm are compared to those of the BNCs learnt in the less informed standard supervised classification paradigm. For the sake of clarity, only the results of the best experimental configuration for each type of BNC are shown for both paradigms. The complete table of results is publicly available in the webpage associated with this paper.

Four different measures —accuracy, recall, precision and F1 [35]— have been used to describe the mean results of the cross validation of the learning process. As can be appreciated in Table 1, the datasets are unbalanced. Thus, evaluating classifiers only by means of accuracy could be unfair. This is confirmed by the experimental results, where the most accurate classifiers are those that always predict the majority (negative) class. Recall, precision and F1 metrics, which provide information on the ability to predict positive examples (implanted embryos), have been also used. In order to fairly analyze these results, the percentage of instances predicted as positive (PPR) is also included. Results of the experiments carried out using only embryonic features as predictors usually show PPR values near 0. They reach, at most, a poor 0.04. Low PPR values could be interesting when the precision of the classifiers is high. However, neither is their performance in terms of precision noteworthy. However, when the training dataset includes predictive features of the cycle, the proportion of predicted positives rises notably, also improving the predictive ability mainly in terms of recall and F1. NB classifiers stand out with the best results in terms of recall (values close to 0.5), precision (values close to 0.3) and F1 (values over 0.3). According to Figure 2, where the decision threshold of our probabilistic classifiers is moved from 0 to 1 in order to build the precision-recall curve, the difference in performance between the classifiers learnt with both datasets is consistent and substantial. The classifiers learnt in the LLP paradigm clearly outperform those learnt in the standard paradigm. With the standard supervised classification paradigm, classifiers learnt using only embryonic features as predictors show no predictive ability. When the combined set of cycle and embryonic predictors is considered, their performance is slightly enhanced (best obtained recall, precision and F1 values are among 0.15 and 0.25).

Using the same layout as Table 2, the results of the second evaluation in the reserved dataset are shown in Table 3. Each row shows the performance in the auxiliary data of an experiment carried out learning a specific type of classifier (NB, TAN or 2DB) with a learning paradigm (standard supervised learning or LLP) using only embryonic features or a combination of embryonic and cycle features as predictive features. Each experiment has been set up using the best performing configuration according to the previous set of experiments (Tab. 2). The list of features selected in each of these experiments, as well as the actual Bayesian network models, are publicly available in the webpage associated with

BNC	Accuracy	Recall	Precision	F1	PPR
NB	0.86 ± 0.00	0.03 ± 0.00	0.40 ± 0.00	0.05 ± 0.00	0.01 ± 0.00
TAN	0.83 ± 0.00	0.04 ± 0.00	0.14 ± 0.01	0.06 ± 0.00	0.04 ± 0.00
2DB	0.84 ± 0.00	0.08 ± 0.00	0.25 ± 0.00	0.12 ± 0.00	0.05 ± 0.00
NB	0.76 ± 0.00	0.49 ± 0.00	0.29 ± 0.00	0.36 ± 0.00	0.23 ± 0.00
TAN	0.79 ± 0.00	0.28 ± 0.01	0.26 ± 0.01	0.27 ± 0.01	0.15 ± 0.00
2DB	0.76 ± 0.00	0.24 ± 0.00	0.20 ± 0.00	0.22 ± 0.00	0.17 ± 0.00

LLP: all available data

BNC	Accuracy	Recall	Precision	F1	PPR
NB	0.86 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
TAN	0.85 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
2DB	0.85 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
NB	0.81 ± 0.00	0.17 ± 0.00	0.24 ± 0.00	0.20 ± 0.00	0.10 ± 0.00
TAN	0.82 ± 0.00	0.10 ± 0.02	0.20 ± 0.03	0.13 ± 0.02	0.07 ± 0.00
2DB	0.81 ± 0.00	0.08 ± 0.00	0.15 ± 0.00	0.11 ± 0.00	0.08 ± 0.00

Standard supervised learning: only labeled data

Table 2: Predictive performance in cross validation of the different BNCs (NB, TAN and 2DB) in terms of accuracy, recall, precision, F1 and predicted positive rate (PPR) metrics. The top table shows the results of our LLP methodology, whereas for the experiments in the bottom table classical supervised BNC learning techniques are used. Each table shows experimental results for two different datasets: in the upper rows, only the embryonic features are used as predictors whereas, in the lower rows, the cycle features are also considered.

this paper.

Performance is again assessed in terms of accuracy, recall, precision, F1 and PPR, calculated as the mean value of 10 repetitions. Comparing the results obtained in both Table 2 and 3, all the classifiers learnt in the different scenarios show a stable generalization performance. In general terms, the different behaviors observed in Table 2 can also be discovered in Table 3. Specifically, the differences between classifiers learnt with the LLP and the standard supervised learning paradigms remain notable. The strengths of the simple NB structure, which prevents this type of models from over fitting the training data, can be observed in the performance increase shown by NB classifiers in the reserved dataset (in terms of recall, precision or F1 metrics).

4 Discussion

The main objective of the assisted reproduction units is the improvement of the pregnancy rate of the ARTs. The proposed ML-based solution deals with two different sets of features to describe the embryos in order to understand the predictive capability of the collected data. Our proposal takes into account all the information collected by physicians, both in terms of examples and features,

BNC	Accuracy	Recall	Precision	F1	PPR
NB	0.81 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
TAN	0.79 ± 0.00	0.02 ± 0.00	0.10 ± 0.00	0.04 ± 0.00	0.04 ± 0.00
2DB	0.78 ± 0.00	0.02 ± 0.00	0.08 ± 0.00	0.03 ± 0.00	0.05 ± 0.00
NB	0.78 ± 0.00	0.69 ± 0.00	0.43 ± 0.00	0.53 ± 0.00	0.28 ± 0.00
TAN	0.74 ± 0.01	0.24 ± 0.01	0.25 ± 0.01	0.24 ± 0.01	0.17 ± 0.00
2DB	0.71 ± 0.00	0.24 ± 0.00	0.22 ± 0.00	0.23 ± 0.00	0.20 ± 0.00

LLP: all available data

BNC	Accuracy	Recall	Precision	F1	PPR
NB	0.82 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
TAN	0.82 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
2DB	0.82 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
NB	0.81 ± 0.00	0.40 ± 0.00	0.47 ± 0.00	0.43 ± 0.00	0.15 ± 0.00
TAN	0.78 ± 0.00	0.13 ± 0.01	0.27 ± 0.02	0.17 ± 0.01	0.08 ± 0.00
2DB	0.78 ± 0.00	0.20 ± 0.00	0.32 ± 0.00	0.25 ± 0.00	0.11 ± 0.00

Standard supervised learning: only labeled data

Table 3: Predictive performance in the reserved dataset of the different BNCs (NB, TAN and 2DB) in terms of accuracy, recall, precision, F1 and predicted positive rate (PPR) metrics. The top table shows the results of our LLP methodology, whereas for the experiments in the bottom table classical supervised BNC learning techniques are used. Each table shows experimental results for two different datasets: in the upper rows, only the embryonic features are used as predictors whereas, in the lower rows, the cycle features are also considered.

to learn classifiers which could be used to improve the implantation (and pregnancy) rates. All the embryos, even those of unknown fate, are efficiently used for training. All the collected features are considered although only a subset of relevant (regarding the class) and non-redundant (among them) features are automatically selected and, finally, used for model learning. Thus, no personal preference determines the allegedly relevant features to be considered.

Learnt classifiers show a limited performance, revealing the difficulty of predicting an embryo implantation based on the data collected in this study. The results in Table 2 vary sharply depending on the set of predictive features used to describe the examples (embryos). On the one hand, the results of the classifiers learnt from the dataset described only by embryonic features are poor: almost no implantation is predicted. This behavior, the prediction of few positive examples, can be interesting if the classifiers are highly precise; i.e., a predictive positive is a real positive example with high probability. However, these classifiers show limited precision. On the other hand, when both embryonic and cycle features are used as predictors, learnt classifiers show a significant improvement in terms of all the metrics. In this case, classifiers achieve a realistic proportion of examples predicted as positive (predicted implantations). In terms of recall, half of the real implanted embryos are predicted as positive. And almost one out of three examples predicted as positive are real implanted embryos (preci-

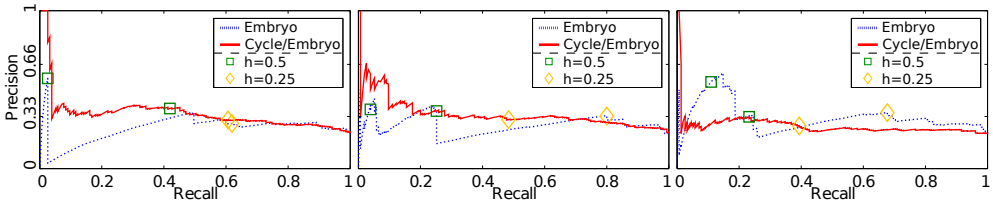


Figure 2: For each BNC type (NB, TAN and 2DB from left to right), precision-recall curve of the best performing classifiers learnt from the dataset with embryonic predictive features (blue line) or with cycle-embryonic features (red line). The highlighted points represent the classifiers using 0.25 and 0.5 as 0/1 class decision threshold, h. An optimistic estimation of the measures is displayed as a completion of the dataset is used for evaluation where the examples with the largest probability according to the classifiers are considered, in each non-full bag, as the real positive examples.

sion). This could be understood as evidence of the low power of the collected (morphological) embryonic features to predict an implantation. Only with the inclusion of features describing the respective cycle does the performance of the learnt classifiers improve. Despite this dramatic difference in the results of Table 2, the classifiers learnt with both sets of features show a similar ability to assign a large variety of different posterior probabilities to the embryos used for evaluation. It can be seen in Figure 2 that, if the threshold of the classifiers learnt only with embryonic features were tuned to optimize a metric (in this case, recall or precision), their performance would match up to that of the classifiers learnt with both cycle and embryonic features. These results remove all doubts about the embryonic features: Their contribution is determinant in the implantation prediction. However, the area under the curve in Figure 2 is significantly larger (specifically in the case of NB and TAN models) when the cycle features are also considered as predictive features. This demonstrates the asserted contribution of the cycle information to the identification of promising embryos for implantation [6, 14, 20, 36].

The performance enhancement derived from the use of our proposal is also notable. The poor performance when classifiers are learnt only with embryonic features is common to both paradigms (LLP and standard supervised learning). The differences are insignificant, with values near to 0 for most of the metrics in both paradigms. Therefore, no conclusion could be fairly drawn from these results. However, with the combination of cycle and embryonic features, the differences are noteworthy, ranging from 0.05 to 0.3 points in terms of recall, precision and F1 metrics. This means that the use of our LLP technique allows classifiers to overcome or even double the performance results obtained with the standard supervised classification paradigm.

The reported differences are confirmed by the second evaluation step, where a reserved validation dataset is used. The results of these experiments, displayed in Table 3, are similar to those reported by the cross validation step. This

behavior demonstrates the generalization ability of the BNCs learnt in this study (with both paradigms). The differences in performance resulting from the implementation of the LLP paradigm remain notable in this second evaluation. The classifiers learnt with the standard paradigm outperform, in some scenarios, those learnt with the proposed approach in terms of precision values. However, with the LLP paradigm, the rate of predicted positive examples is consistently doubled, promoting competitive classifiers in terms of recall (almost a difference of 0.3 points).

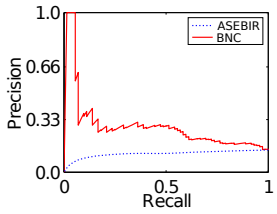
In order to build the datasets where embryos are described by a combination of embryonic and cycle features, the values of the cycle features are replicated for all the embryos of the same cycle. This unquestionably breaks the independence and identically distributed (i.i.d.) assumption. The relational nature of this problem, where a cycle is related to one or more embryos, could be handled by specific techniques, such as probabilistic relational models [37]. However, the conclusion drawn in this study is still valid. The use of cycle features enhances the performance of the learnt classifiers.

To the best of our knowledge, the use of relational learning techniques would be also novel in the ART literature. However, its application is not straightforward. An adaptation of the current relational learning techniques to the problem of learning from label proportions should be necessary in order to take full advantage of the available information. Otherwise, a poor solution which discards embryos of unknown fate is straightforward.

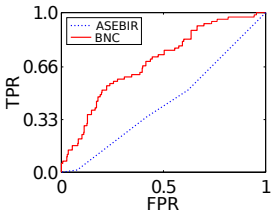
Although promising classification models have been learnt, these moderate results can be explained by the apparent limited predictive capability of the collected features. Observing the list of features selected by the proposed technique in different experiments (the complete list is publicly available in the webpage associated with this work), a subset of relevant features is constantly selected (*SER*, *PB* and, mainly, *nCel.2* and *frag.2*). Although others are regularly selected (e.g., *PVS*, *vac.2*, *symmet.2* or *symmet.2*), the number of embryonic features eventually selected ranges from 4 to 7 (out of the original 14 features). This severe reduction would be in agreement with several previous studies which claim that the selection of embryos based exclusively on morphological factors is not efficient [6, 7, 16]. In this study, the features collected for the oocytes/embryos, those recommended by the ASEBIR protocol [28], are a set of morphological features. The search, study and collection of other non-morphological features, such as pre-implantation genetic diagnosis, embryo metabolomic and proteomic analysis, embryo morphokinetics analysis or endometrial receptivity tests, have been proposed in the related literature [18, 38, 39, 40, 41, 42, 43, 44, 45]. This research line will surely allow authors to progress in the answer of this open question; its solution is expected to bring a significant leap forward in the ability to predict the embryo implantation and, consequently, the ART success.

Based on the results obtained, it can be asserted that a recommender system for the embryo selection problem based on the LP-learnt model would provide valuable information that could imply an improvement in the selection of promising embryos. As shown in Figure 3(a), ordering embryos according to

(a) List of all the embryos ordered in the horizontal axis according to their probability of implantation, and in the vertical axis depending on their ASEBIR grade.



(b) PR curve using embryos of known fate.



(c) ROC curve using embryos of known fate.

Figure 3: Graphical comparison of our classification model and the ASEBIR grade [28].

their implantation probability does not completely match up with the ASEBIR ranking. ASEBIR proposes an ordinal four-categories ranking (A, B, C and D), where A and D respectively indicate the best and worst embryos. In detail, our classifiers usually agree with ASEBIR's criteria on the embryos identified as top-quality (top-left corner in Fig. 3(a)). However, numerous quality C and D embryos are considered by our classification model as more promising than a substantial set of quality B embryos. Both the precision-recall (PR, Fig. 3(b)) and the receiver operator characteristic (ROC, Fig. 3(c)) curves graphically show the enhanced performance of our model with respect to the ASEBIR ranking. The observed disagreement in medium-quality embryos was not surprising since, as has been previously reported [9, 16], the most difficult task is not the identification of the highly promising embryos, but the classification of those of medium-quality. The reordering suggested by our ML models is supported by the daily practice of our group of physicians, who are already considering an analogous variation of the ASEBIR criteria based on their direct observation of the evaluated embryos.

5 Conclusions

In this paper, the problem of implantation prediction of the ART problem is analyzed by means of novel ML techniques. The proposed solution learns classification models (BNCs) taking full advantage of all the available weakly super-

vised data. This technique has been used to study the data of a set of embryos gathered in the laboratory.

Our solution automatically selects the relevant features to be considered for model learning, initially considering all the information collected during the whole ART procedure, and discarding no embryo independently of the information available about its implantation. This can be carried out without the intervention of the physician. The results of this study are in line with the hypothesis of a combination of embryonic and cycle aspects determining the implantation. According to the results, the data collected for this study cannot fully describe an embryo implantation, although the inclusion of the cycle features enhances the classification performance. Moreover, the use of our comprehensive solution takes full advantage of the available information. In this way, a significantly enhanced performance is reported with respect to a solution based on the standard supervised classification paradigm. Obtained classifiers have been proved to rank the medium-quality embryos of this study more consistently than ASEBIR grade. In this way, the probabilistic assessment of the classifiers could be consistently used for embryo quality grading.

Acknowledgments

This work has been partially supported by the Basque Government (IT609-13, Elkartek BID3A), the Spanish Ministry of Economy and Competitiveness (TIN2013-41272-P) and the University-Society Project 15/19 (Basque Government and University of the Basque Country UPV/EHU). Jerónimo Hernández-González is supported by a post-doc grant (UPV/EHU). Jose A. Lozano is also supported by BERC program 2014-2017 (Basque Government) and Severo Ochoa Program SEV-2013-0323 (Spanish Ministry of Economy and Competitiveness).

References

[1] L. Engmann, N. Maconochie, S. L. 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. *Hum. Reprod.*, 16(12):2598–2605, 2001.

[2] P. M. Martin and H. G. Welch. Probabilities for singleton and multiple pregnancies after in vitro fertilization. *Fertil. Steril.*, 70(3):478–481, 1998.

[3] N. Gleicher and D. Barad. The relative myth of elective single embryo transfer. *Hum. Reprod.*, 21(6):1337–1344, 2006.

[4] F. Lesourd, O. Parant, M. Clouet-Delannoy, and J. Parinaud. Clinical and biological parameters influencing implantation: score to determine number of embryos to transfer. *Reprod. Biomed. Online*, 12(4):453–459, 2006.

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

[6] H. Achache and A. Revel. Endometrial receptivity markers, the journey to successful embryo implantation. *Hum. Reprod. Update*, 12(6):731–746, 2006.

[7] T. Ebner, M. Moser, M. Sommergruber, and G. Tews. Selection based on morphological assessment of oocytes and embryos at different stages of preimplantation development: a review. *Hum. Reprod. Update*, 9(3):251–262, 2003.

[8] J. M. Cummins, T. M. Breen, K. L. Harrison, J. M. Shaw, L. M. Wilson, and J. F. Hennessey. A formula for scoring human embryo growth rates in in vitro fertilization: its value in predicting pregnancy and in comparison with visual estimates of embryo quality. *J. In Vitro Fert. Embryo Transf.*, 3(5):284–295, 1986.

[9] J. D. Fisch, H. Rodriguez, R. Ross, G. Overby, and G. Sher. The graduated embryo score (GES) predicts blastocyst formation and pregnancy rate from cleavage-stage embryos. *Hum. Reprod.*, 16(9):1970–1975, 2001.

[10] R. R. Saith, A. Srinivasan, D. Michie, and I. L. Sargent. Relationships between the developmental potential of human in-vitro fertilization embryos and features describing the embryo, oocyte and follicle. *Hum. Reprod. Update*, 4(2):121–134, 1998.

[11] C. V. Steer, C. L. Mills, S. L. Tan, S. Campbell, and R. G. Edwards. The cumulative embryo score: a predictive embryo scoring technique to select the optimal number of embryos to transfer in an in-vitro fertilization and embryo transfer programme. *Hum. Reprod.*, 7(1):117–119, 1992.

[12] E. Van Royen, K. Mangelschots, D. De Neubourg, M. Valkenburg, M. Van de Meerssche, G. Ryckaert, W. Eestermans, and J. Gerris. Characterization of a top quality embryo, a step towards single-embryo transfer. *Hum. Reprod.*, 14(9):2345–2349, 1999.

[13] S. Ziebe, K. Petersen, S. Lindenberg, A. G. Andersen, A. Gabrielsen, and A. N. Andersen. Embryo morphology or cleavage stage: how to select the best embryos for transfer after in-vitro fertilization. *Hum. Reprod.*, 12(7):1545–1549, 1997.

[14] G. Corani, C. Magli, A. Giusti, L. Gianaroli, and L. M. Gambardella. A Bayesian network model for predicting pregnancy after in vitro fertilization. *Comput. Biol. Med.*, 43(11):1783–1792, 2013.

[15] A. Debón, I. Molina, S. Cabrera, and A. Pellicer. Mathematical methodology to obtain and compare different embryo scores. *Math. Comput. Model.*, 57(5-6):1380–1394, 2013.

[16] F. Guerif, A. Le Gouge, B. Giraudeau, J. Poindron, R. Bidault, O. Gasnier, and D. Royere. 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):1973–1981, 2007.

[17] D. A. Morales, E. Bengoetxea, and P. Larrañaga. Selection of human embryos for transfer by Bayesian classifiers. *Comput. Biol. Med.*, 38(11–12):1177–1186, 2008.

[18] G. Patrizi, C. Manna, C. Moscatelli, and L. Nieddu. Pattern recognition methods in human-assisted reproduction. *Int. Trans. Oper. Res.*, 11(4):365–379, 2004.

[19] C. Racowsky, L. Ohno-Machado, J. Kim, and J. D. Biggers. Is there an advantage in scoring early embryos on more than one day? *Hum. Reprod.*, 24(9):2104–2113, 2009.

[20] S. A. Roberts. Models for assisted conception data with embryo-specific covariates. *Stat. Med.*, 26(1):156–170, 2007.

[21] J. Holte, L. Berglund, K. Milton, C. Garello, G. Gennarelli, A. Revelli, and T. Bergh. Construction of an evidence-based integrated morphology cleavage embryo score for implantation potential of embryos scored and transferred on day 2 after oocyte retrieval. *Hum. Reprod.*, 22(2):548–557, 2007.

[22] J. Hernández-González, I. Inza, and J. A. Lozano. Learning Bayesian network classifiers from label proportions. *Pattern Recognit.*, 46(12):3425–3440, 2013.

[23] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *J. Mach. Learn. Res.*, 10:2349–2374, 2009.

[24] C. Bielza and P. Larrañaga. Discrete Bayesian network classifiers: a survey. *ACM Comput. Surv.*, 47(1):5, 2014.

[25] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, 1997.

[26] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[27] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[28] M. Ardoy and G. Calderón. *Clinical Embryology Papers: ASEBIR criteria for the morphological evaluation of human oocytes, early embryos and blastocysts*. Asociación para el Estudio de la Biología de la Reproducción (ASEBIR), Madrid, Spain, 2nd edition, 2008.

[29] D. J. Hand and K. Yu. Idiot's Bayes—not so stupid after all? *Int. Stat. Rev.*, 69(3):385–398, 2001.

[30] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 29(2–3):131–163, 1997.

[31] M. Sahami. Learning limited dependence Bayesian classifiers. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 335–338, 1996.

[32] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Learning in Graphical Models, 1995.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 39(1):1–38, 1977.

[34] M. A. Hall. *Correlation-based feature selection for machine learning.* PhD thesis, Department of Computer Science, The University of Waikato, 1999.

[35] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.*, 45(4):427–437, 2009.

[36] A. L. Speirs, A. Lopata, M. J. Gronow, G. N. Kellow, and W. I.H. Johnston. Analysis of the benefits and risks of multiple embryo transfer. *Fertil. Steril.*, 39(4):468–471, 1983.

[37] L. Getoor, N. Friedman, D. Koller, A. Pfeffer, and B. Taskar. Probabilistic relational models. In *Introduction to statistical relational learning*, chapter 5, pages 175–200. MIT press, 2007.

[38] D. R. Brison, F. D. Houghton, D. Falconer, S. A. Roberts, J. Hawkhead, P. G. Humpherson, B. A. Lieberman, and H. J. Leese. Identification of viable embryos in IVF by non-invasive measurement of amino acid turnover. *Hum. Reprod.*, 19(10):2319–2324, 2004.

[39] D. R. Brison, K. Hollywood, R. Arnesen, and R. Goodacre. Predicting human embryo viability: the road to non-invasive analysis of the secretome using metabolic footprinting. *Reprod. Biomed. Online*, 15(3):296–302, 2007.

[40] D. K. Gardner, M. Lane, J. Stevens, and W. B. Schoolcraft. Noninvasive assessment of human embryo nutrient consumption as a measure of developmental potential. *Fertil. Steril.*, 76(6):1175–1180, 2001.

[41] M. G. Katz-Jaffe, D. K. Gardner, and W. B. Schoolcraft. Proteomic analysis of individual human embryos to identify novel biomarkers of development and viability. *Fertil. Steril.*, 85(1):101–107, 2006.

[42] C. Mendoza, E. Ruiz-Requena, E. Ortega, N. Cremades, F. Martinez, R. Bernabeu, E. Greco, and J. Tesarik. Follicular fluid markers of oocyte developmental potential. *Hum. Reprod.*, 17(4):1017–1022, 2002.

[43] W. E. Roudebush, J. D. Wininger, A. E. Jones, G. Wright, A. A. Toledo, H. I. Kort, J. B. Massey, and D. B. Shapiro. Embryonic platelet-activating factor: an indicator of embryo viability. *Hum. Reprod.*, 17(5):1306–1310, 2002.

[44] E. Seli, C. G. Vergouw, H. Morita, L. Botros, P. Roos, C. B. Lambalk, N. Yamashita, O. Kato, and D. Sakkas. Noninvasive metabolomic profiling as an adjunct to morphology for noninvasive embryo assessment in women undergoing single embryo transfer. *Fertil. Steril.*, 94(2):535–542, 2010.

[45] G. Sher, L. Keskintepe, M. Nouriani, R. Roussev, and J. Batzofin. Expression of sHLA-G in supernatants of individually cultured 46-h embryos: a potentially valuable indicator of 'embryo competency' and IVF outcome. *Reprod. Biomed. Online*, 9(1):74–78, 2004.