UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

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

---

Author:
Adrián TORRES MARTÍN

Supervisors:
Jerónimo HERNÁNDEZ-GONZÁLEZ
Jesús CERQUIDES

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

July 1, 2021

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

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

by Adrián TORRES MARTÍN

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. We assemble an experimental setup consisting of alternative and simpler methods as a basic reference point to compare against. They are built in an incremental way in order to test different aspects of our probabilistic graphical model. We show the benefits of using an EM algorithm and the importance of the cycles characteristics. Special attention is given to the relation between the models 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 infer that ASEBIR score provides a good summary of morphological features.

Acknowledgements

First of all, I would like to express my sincere gratitude to the supervisors of my Master's thesis, Jerónimo Hernández-González and Jesús Cerquides for their guidance and advice over the development of the project. Their continuous supervision helped me to keep the development up to date and our discussions allowed me to understand and appreciate a very interesting topic. This project would not have been possible without their valuable insights and suggestions.

I would also like to thank my parents for their unconditional support throughout all my studies. Without them these years would have been much harder.

Contents

Chapter 1

Embryo selection for IVF

1.1 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 in order 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. 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 (Engmann et al., 2001), but this leads to higher multiple-birth rates, which is considered risky for both mother and the developing fetuses (Engmann et al., 2001; Report, 2001). 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 (Cuevas-Sáiz et al., 2018). A unified criteria was created to address this problem: the ASEBIR protocol (Ardoy and Calderon, 2008). 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 (Corani et al., 2013; Guérif et al., 2007; Hernández-González et al., 2018; Kragh et al., 2019). 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 (Hernández-González et al., 2018) try to benefit from cycles with partial implantation (not all the transferred embryos were implanted).

In this work we consider the model proposed by Valls Murcia (2021), which expands on the idea of using partial implantation. The presented probabilistic graphical model works under the assumption of independence between embryos and cycles, and accounts for a third source of uncertainty corresponding to unknown factors. An EM algorithm is used to learn the hyperparameters in the context of partially observed data and latent variables.

Our main goal is to perform a thorough experimental validation of the model using real data. In our experimental setting, we compare our model with a set of alternative baseline models which were created in an incremental way in order to test the different properties of the model. A special part of our attention is devoted to the ASEBIR score and its relationships with our model and the alternative baselines. This quality grade is given as a feature in the dataset and we test whether our model is able to learn from the available data or just relies on this key feature.

The performance of the models is tested using suitable metrics, which is a non-trivial task because of the partially observed nature of the problem. We use measures that work in different evaluation settings and that may assess different benefits of the model. In particular, AUC-ROC is evaluated using only embryos with known outcome, which provides a measure of the predictive power of the model for these specific cases. We also compute the negative log-likelihood which is appropriate to consider partially implanted cycles. All the code used in this project can be found in a GitHub repository.

The document is organized as follows. Next, we present the state of the art. In Chapter 2 we describe the dataset of our case study. In Chapter 3, the model is presented as well as the learning algorithm. In Chapter 4 the experimental setup is explained, introducing the different probabilistic classifiers, the baseline methods to compare against and the used metrics. Then, their results are shown and discussed in Chapter 5. Finally, in Chapter 6 conclusions are drawn and a few open lines for future work are presented.

1.2 State of the art

ART techniques and, in particular, IVF treatments present complex processes involving a large number of variables, providing an ideal context for the application of machine learning and artificial intelligence techniques. Since the popularization of infertility treatments, there have been many approaches to the problem of predicting the outcomes of the procedures and the selection of the most promising embryos (Siristatidis et al., 2011; Fernandez et al., 2020). The field has advanced in parallel to the methods and computational innovations: from classical statistical techniques to more complex ML techniques such as Bayesian Networks (Morales et al., 2008; Corani et al., 2013; Hernández-González et al., 2018), Support Vector Machines (Uyar, Bener, and Ciray, 2014) and recently deep learning methods (Kragh et al., 2019).

The traditional approach to predict implantation uses morphological characteristics of embryos and clinical information about the treatment, female patient and male patient. Many techniques have been used to model this data and improve success rates. In Morales et al. (2008) embryo selection is addressed using different Bayesian classifiers with diverse subsets of variables. They find that the most significant subset of variables is indeed the one used by embryologists in normal practice.

Most of the techniques applied to ART use supervised classification, where the models learn from previous labeled examples. However, current medical techniques are able to know the number of embryos predicted in a cycle, not their identity. This presents a problem for the usual supervised framework. In Morales et al. (2008) this issue is addressed joining all embryos in a cycle as a single instance and dealing with the problem at a batch level. However, the corresponding classes are just 1 or 0 depending on whether there was an implantation on the cycle. Much information is lost regarding the number of embryos implanted. In other cases the embryos with

unknown outcome are completely disregarded (Debón et al., 2013; Racowsky et al., 2009; Kragh et al., 2019).

In Hernández-González et al. (2018) a set of Bayesian Networks is proposed as probabilistic classifiers taking full advantage of the weakly supervised data. The models learn not only from the embryos with known outcome but also from those in partially implanted cycles through their label proportions.

Another widespread approach is the embryo-uterine model (EU), introduced by Speirs et al. (1983) and later extended by Zhou and Weinberg (1998), which assumes that, for a pregnancy to happen both a receptive uterus and viable embryo are necessary. This model is formed by two separate components (embryo [E] and uterus [U]) and the predicted probability of implantation is the product of both outputs. Moreover, the model is also compatible with multiple transferred embryos, assuming independence between them. However, this approach faces once again the problem of partial observability. In this context even for cases where no embryo implanted, we have unknown latent variables. If that is the case, we do not known if the embryos were not viable, if the uterus was not receptive or both. In Roberts (2007) this is addressed via the Expectation-Maximization (EM) algorithm. In Corani et al. (2013) a Bayesian network is trained with an averaging approach as an alternative to MAP estimation. In this case the set of variables used for both embryos and cycles is very reduced.

In Valls Murcia (2021). an extended probabilistic graphical model is presented. This model also deals with cycles with partial implantation, under the assumption of independence between embryos and cycles. Moreover it takes into account a third source of uncertainty corresponding to external or unknown factors. To deal with the appearance of latent variables, it employs an EM algorithm as learning method.

The other completely different approach to the problem consists in obtaining the embryo characteristics directly from images. In Patrizi et al. (2004) a pattern recognition algorithm is presented which is able to classify embryos into a number of classes. This procedure was recognized (Manna et al., 2004) to obtain better results than the judgement of experts.

Recently, the advance of computational processing has enabled Deep Learning techniques, which are able to explore complex nonlinear patterns and extract high-level features, hence its importance in image analysis. In particular, convolutional neural networks have been used in ART, analyzing not only static but time-lapse imaging. In Kragh et al. (2019) a deep learning method is proposed which is able to predict inner cell mass (ICM) and trophectoderm grades (TE) with a convolutional neural network. Moreover a recurrent neural network is applied on top of that to account for the temporal information provided by the multiple frames obtained during the whole process in which embryos are cultured.

Chapter 2

Data

The database, originally studied in Hernández-González et al. (2018), 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, only some of which were actually transferred (see Figure 2.1). Cycles are described by 25 features related to the female patient, the sperm donor, the stimulation procedure and summary attributes about associated embryos. Embryos are described by 20 features, out of which 13 summaryze different morphological characteristics of different stages of development (up to 48 hours after fertilization, when transference was carried out).

2.1 Data Exploration

The main problem when dealing with real data from ART is that we face partially observed data. Since we are not able to know the identity of the implanted embryos in multiple transfers, we do not know their actual outcome; in many cases we only know the proportion of embryos implanted in the corresponding cycle.

Out of the 604 cycles, 192 resulted in a pregnancy with 253 embryos implanted. Of these successful cycles, in 57 of them all the transferred 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 cycles with not all embryos implanted), and for the rest, 1871 embryos, we do not have any information (not transferred embryos). These counts can be seen in Figure 2.2.

In this work we devote a large portion of our attention to the grade given by the ASEBIR protocol to each embryo and how the models interact with it. This grading

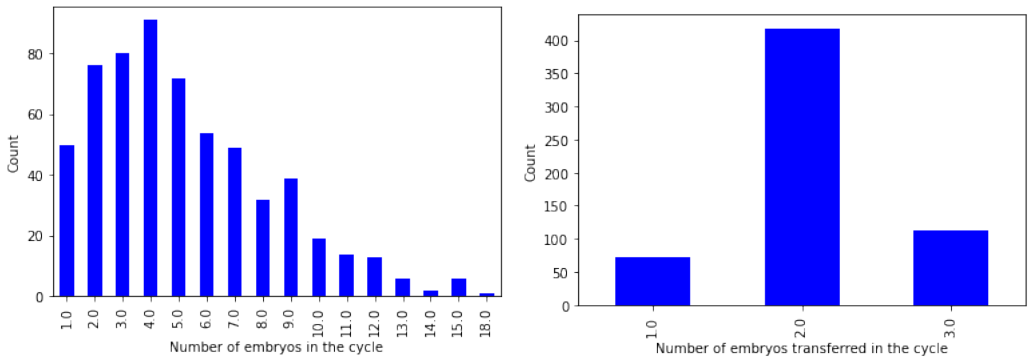

FIGURE 2.1: Distribution of cycles depending on the number of associated embryos (left) and transferred embryos (right).

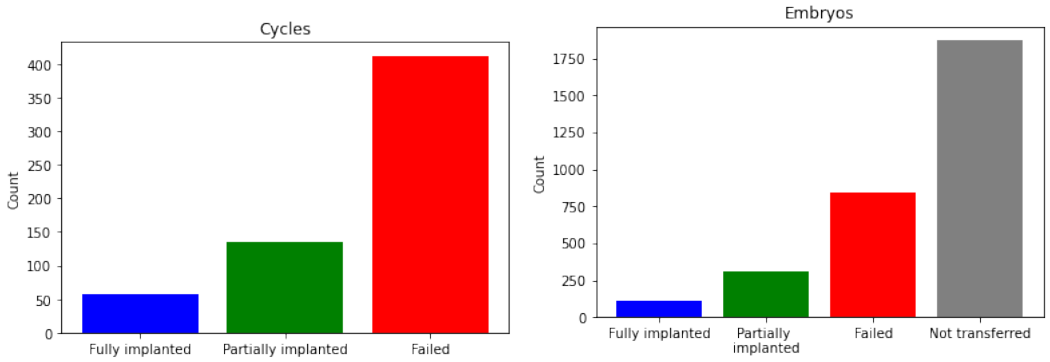

FIGURE 2.2: Number of cycles depending on their outcome. Similarly, for embryos: implanted, transferred but unknown (in partially implanted cycle), not implanted or not transferred.

system assigns to each embryo a category based on its morphological characteristics. The grading scale has four different categories ranging from A (the most promising embryos) to D (the embryos with the poorest quality). Figure 2.3 shows that this quality score is a decisive factor in the selection process performed by embryologists.

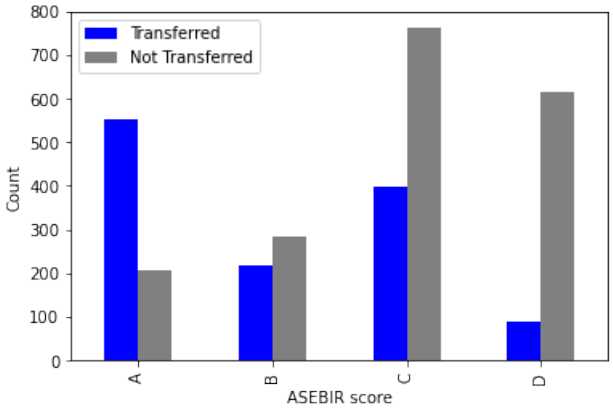

FIGURE 2.3: Number of embryos transferred or not for each ASEBIR grade.

Ideally, there should be a clear difference in the implantation rates between the different graded embryos. Naturally, this data is only available for transferred embryos and, as it is intended, there are many more good quality embryos transferred than those of poor quality. To provide a clear comparison between the different classes we display in Figure 2.4 the fraction of transferred embryos of each quality that had each cycle outcome: fully implanted, partially implanted (we do not know if the considered embryo was indeed implanted) and not implanted. The implantation rates are also shown in Table 2.1 to have a clearer representation of the slight differences.

The most noticeable feature of the implantation rates is the fact that none of the D quality embryos were unequivocally implanted, although some could have been in partially implanted cycles. Moreover, the proportion of embryos not implanted is considerably higher than for better quality scores. This fact, along with the high proportion of implanted embryos with A quality, are positive indicators regarding the effectiveness of the ASEBIR protocol. Lastly, there does not seem to be a significant improvement in implantation rates from C to B quality embryos.

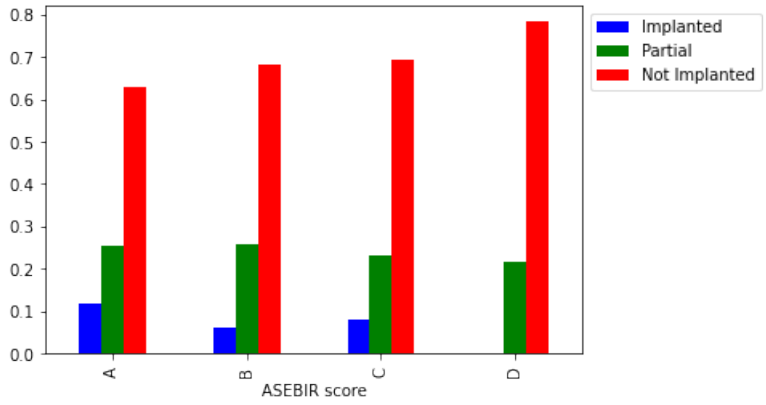

FIGURE 2.4: Fraction of transferred embryos with each outcome for each ASEBIR score.

ASEBIR Score	Implanted	Partial	Not Implanted
A	0.12	0.25	0.63
B	0.06	0.26	0.68
C	0.08	0.23	0.69
D	0.00	0.22	0.78

TABLE 2.1: Fraction of transferred embryos with each outcome for each ASEBIR score.

2.2 Features and preprocessing

The full set of collected features for each ART cycle is shown in Figure 2.2. These features describe characteristics of different components of the cycle: the female patient, the male patient and the stimulation procedure. Moreover, a general summary of the associated embryos is provided.

Similarly, Figure 2.3 shows the features collected for each individual embryo. They are mainly morphological characteristics at different stages of embryo development. The outcome of the embryo is described by three variables. *Transfer* represents whether the embryo was selected to transfer. *Vitrificado* represents whether the embryo was frozen because of a surplus of embryos in the cycle. If the embryo was indeed transferred then implantation is partially encoded in the *TasaExito* variable from its associated cycle. Only where this variable is 0 (no embryos implanted in the cycle) or 1 (all embryos implanted) we know unequivocally the outcome of the embryo. This is due to the aforementioned inability to identify implanted embryos.

The first step in the preprocessing of the data was to correct some mistakes in the summary of the embryos of certain cycles. Some cycles displayed a number of associated or transferred embryos that did not coincide with the true values.

Both datasets have many categorical features alongside numerical ones. Some probabilistic classifiers are able to work directly with heterogeneous data but we want to provide a standardized dataset so all methods described in the experimental setup (Chapter 4) work under the same conditions. Therefore, all binary categorical variables are transformed into numerical features. Then, for all multi-categorical variables we use a one-hot encoding strategy, creating new variables for each possible category.

To deal with the relation between cycles and embryos we created general variables that record which embryos belong to each cycle and which were transferred.

TABLE 2.2: Features collected for each ART cycle.

Variable	Possible values	Description
Codigo	Numeric	Identifier of the cycle
TEsteril	Numeric	Time since infertility was detected
Indicac	endometriosis, fracasoia, tubarico, masculino, mixto, desconocido	Indication of the cycle
Features related to female patient		
Edad	Numeric	Age
IMC	Numeric	Body mass index
EmbPrev	No, Yes	Has she ever got pregnant?
AboPrev	No, Yes	Has she ever aborted?
FSH	Numeric	Quantity of follicle-stimulating hormone
CiclosPrevios	Numeric	Number of previously undergone cycles
AMH	Numeric	Quantity of anti-mullerian hormone
folAntral	Numeric	Nomber of antral follicles
E2	Numeric	Quantity of estradiol
P4	Numeric	Quantity of progesterone
lEnd	Numeric	Endometrial thickness
Features related to male patient		
caSemen	A, N, O, OA, OAT	Quality of the semen
REM	Numeric	Total pregressive sperm recovery
Features related to stimulation		
Protocol	PC, PL	Stimulation protocol
Estimul	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		
nEmbObten	Numeric	Number of embryos
TasaFertil	Numeric	nEmbObten / Number of mature oocytes (MII state)
nEmbTrans	Numeric	Number of transferred embryos
Outcome		
TasaExito	Numeric	Number of implanted embryos/ nEmbTrans

TABLE 2.3: Features collected for each individual embryo.

Variable	Possible values	Description
CodigoCiclo	Numeric	Identifier of the cycle
CodigoOvoc	Numeric	Identifier of the embryo
Tecnica	IVF, ICSI	Fertilization technique
Features related to oocytes		
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 D+1		
CP.1	Numeric	Number of polar bodies
Z	Z1, Z2, Z3, Z4	Scott's pronuclear grade
Features at D+2		
nCel+2	Numeric	Number of cells
frag+2	Numeric	Percentage of cell fragmentation
simet+2	No, Yes	Are the blastomeres symmetric?
ZP+2	Normal, Abnormal	Zona pellucida
vac+2	No, Few, Many	Presence of vacuoles
multiNuc+2	No, Yes	Presence of multi-nucleation in a cell
CALIDAD+2	A, B, C, D	ASEBIR quality grade
Outcome		
Transfer	No, Yes	Embryo selected for transference
Vitrificado	No, Surplus	Surplus' embryos to froze
TasaExito	Numeric	TasaExito of the associated cycle

Moreover, they also encode implantation information. These variables are only used for internal work when creating the model and executing the learning algorithm. Of course, they are not actually provided to the probabilistic classifiers.

Then we can remove all variables from the datasets that record either identifiers or outcomes. In the cycles dataset we also remove the *nEmbTrans* variable (because it is already encoded in the internal variables) and the *AMH* variable (because it has many missing values).

After all this process we are left with 36 features for cycles and 25 for embryos. Both datasets are then standardized (centered and scaled to unit variance) and ready to be fed to the probabilistic classifiers.

Chapter 3

Method

In this work we employ a probabilistic graphical model originally presented in Valls Murcia (2021) that uses the available information from both cycles and individual embryos, and considers a third source of uncertainty related with unknown factors (Coughlan et al., 2015).

3.1 Introduction to PGM

Probabilistic Graphical Models (PGM) is a framework that provides structure to represent and manipulate complex joint distributions in a compact way. General probabilistic models may face complex systems where the number of random variables is too large to compute a joint distribution if no assumption is made (the complexity grows exponentially with the number of variables). PGM combines the knowledge from probability theory and graph theory. They use a graph-based representation as the basic structure to encode all the probability distributions.

This type of structure allows for an explicit representation of the domain knowledge to be directly applied on the model. The framework provides a really efficient way to perform inference. Many probability based operations (marginalization, conditioning, belief propagation, etc.) are prepared to work directly on the graph structure and are generally much faster than manipulating the joint distribution directly. Regarding the learning process, PGMs are able to use the provided data to not only learn the parameters of the distributions but also construct the model according to the connections found between variables. It is also relevant to our purpose to point out that often the learning algorithms use inference as a recurrent part of their process (e.g., EM algorithm).

There are two main types of structures to represent the probability distribution:

- Directed acyclic graphs (DAG): Bayesian Networks. The graph represents the set of conditional independence assumptions (edges) over the several random variables (nodes). The associated parameters are the conditional probability distributions needed to obtain the joint distribution.

- Undirected graphs: Markov Networks. The graph provides a skeleton (based on independence assumptions) for factorizing a distribution.

Each structure may be useful depending on the problem at hand. Markov Networks can be applied to express certain dependencies that Bayesian Networks cannot. However Bayesian Networks are usually easier to interpret (because of directionality) and do not need to compute a normalization term.

In this work we focus on the models provided by Bayesian Network. Let us define first the concept of DAG.

Definition. Let $V = \{1, \ldots, n\}$ be a set of vertices and $E = \{(u, v) : u, v \in V,\ u \neq v\}$ a set of edges between those vertices. A directed acyclic graph (DAG) G is a pair (V, E), where there are no directed cycles. This is, there are no directed sequences of edges where the starting vertex of the first edge equals the ending vertex of its last edge.

Then a Bayesian Network is defined in the following way.

Definition. A Bayesian Network is formed by a DAG $G = (V, E)$ and a set of parameters $\Theta = (\theta_1, \ldots, \theta_n)$. Each vertex $i \in V$ is associated with a random variable X_i and its corresponding conditional probability distribution is $p(x_i|\mathbf{pa}_i; \theta_i)$, where $\mathbf{pa}_i$ is the set of vertices with an edge towards X_i.

The complete joint probability distribution of a Bayesian Network M is the product of the conditional probabilities given by the graph, which in fact is just a simplification of the chain rule:

$$p_M(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i; \theta_i) \tag{3.1}$$

Example. Consider the DAG in Figure 3.1. The corresponding Bayesian Network's decomposition is:

$$p_M(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_2)p(x_6|x_4, x_5) \tag{3.2}$$

Note that this expression is much simpler than the one obtained using directly the chain rule on the joint probability distribution. This DAG encodes many conditional independence assumptions, such as $X_6 \perp\!\!\!\perp X_2 \mid X_4, X_5$ or $X_3 \perp\!\!\!\perp X_2 \mid X_1$.

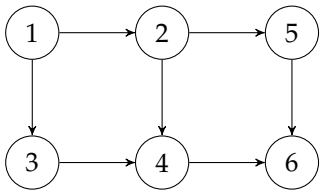

FIGURE 3.1: Representation of a DAG with 6 nodes.

In Bayesian Networks, when a variable or group of variables repeat several times a plate model representation is often used to simplify the graphs. That is, if there is a group of variables that repeat together sharing the same set of attributes and under the same probabilistic model for each repetition, then we use a plate to represent them altogether. This is particularly useful in problems with recurring structures such as temporal scenarios or language models (e.g, Latent Dirichlet Allocation).

Example. Consider a class with N students and S subjects. Each subject has a specific difficulty D_s and each student has a specific intelligence I_n. Both difficulty of the subject and intelligence determine the grade G_{sn} of the corresponding student. We assume that the intelligence is general and is independent from the subject. Then, the Bayesian Network is represented by the plate model in Figure 3.2.

3.2 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

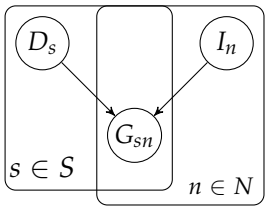

FIGURE 3.2: Example of a plate model assuming general intelligence.

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 (Coughlan et al., 2015) 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 be 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

$$p(w_e^c \mid x_e^c; \alpha), \tag{3.3}$$

where α is the hyperparameter of such distribution.

Similarly, let v_c be 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

$$p(r_c \mid v_c; \beta), \tag{3.4}$$

where β is the hyperparameter of such distribution. Both w_e^c and r_c are modeled using probabilistic classifiers. Then we use the output predicted probabilities for each class as $p(w_e^c \mid x_e^c; \alpha)$ and $p(r_c \mid v_c; \beta)$.

Let s_e^c be 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

$$i_e^c \sim \text{Bernoulli}(\theta_{w_e^c \cdot r_c \cdot s_e^c}), \tag{3.5}$$

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 be 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),

$$y_c = \sum_{e \in E_c} i_e^c, \tag{3.6}$$

where E_c is the set of embryos associated to cycle c.

Figure 3.3 shows the complete graphical representation of the model. The shadowed variables are the observed ones (features, embryo selection 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

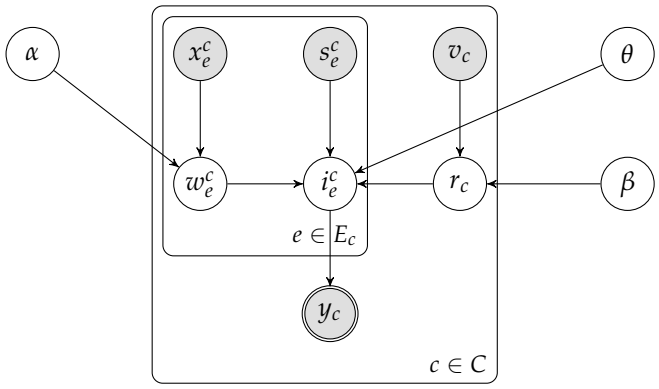

FIGURE 3.3: Graphical description of the proposed model. Shadowed nodes represent observed variables. Double line denotes a deterministic variable.

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

$$p(\mathbf{x}, \mathbf{w}, \mathbf{v}, \mathbf{r}, \mathbf{s}, \mathbf{i}, \mathbf{y}; \alpha, \beta, \theta) = p(\mathbf{w}|\mathbf{x}; \alpha)p(\mathbf{x})p(\mathbf{r}|\mathbf{v}; \beta)p(\mathbf{v})p(\mathbf{s})p(\mathbf{y}|\mathbf{i})p(\mathbf{i}|\mathbf{w}, \mathbf{r}, \mathbf{s}; \theta) \tag{3.7}$$

Since the model assumes independence among instances given the characteristics of the cycles and embryos we can decompose the probability even further:

$$p(\mathbf{w}|\mathbf{x}, \alpha) = \prod_{c=1}^{B} \prod_{e \in E_c} p(w_e^c \mid x_e^c; \alpha) \tag{3.8}$$

$$p(\mathbf{r}|\mathbf{v}, \beta) = \prod_{c=1}^{B} p(r_c \mid v_c; \beta), \tag{3.9}$$

where B is the total number of cycles.

The relation between $\mathbf{y}$ and $\mathbf{i}$ is actually deterministic. We can reduce the probabilities to only terms where $\mathbf{i}$ is compatible with the selections $\{s_e^c\}$ and the known outcomes $\{y_c\}$. This set of vectors is denoted by $\mathbb{I}_{\mathbf{s},\mathbf{y}}$. 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 $[0, 0, 1, 0]$. If $\mathbf{i} \notin \mathbb{I}_{\mathbf{s},\mathbf{y}}$ then $p(\mathbf{y}|\mathbf{i})p(\mathbf{i}|\mathbf{w}, \mathbf{r}, \mathbf{s}; \theta) = 0$.

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

$$p(\mathbf{y}|\mathbf{x}, \mathbf{v}, \mathbf{s}; \alpha, \beta, \theta) = \sum_{\mathbf{r}} p(\mathbf{r}|\mathbf{v}; \beta) \sum_{\check{\imath} \in \mathbb{I}_{\mathbf{s},\mathbf{y}}} \sum_{\mathbf{w}} p(\check{\mathbf{i}}|\mathbf{w}, \mathbf{r}, \mathbf{s}; \theta)p(\mathbf{w}|\mathbf{x}; \alpha) \tag{3.10}$$

3.2.1 An EM algorithm to learn the parameters of our model

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 (Dempster,

Laird, and Rubin, 1977) 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.

In a general setting, EM algorithms aim to find maximum likelihood estimations of parameters when unobserved variables are present. Let X be the observed variables in the model and Z the unobserved latent ones. The complete log-likelihood is $l(\theta; X, Z)$, where θ are the parameters which we want to estimate maximizing the likelihood.

The **E-step** consists in computing the conditional expected value of the log-likelihood given the observed variables and the current estimations of the parameters θ_t:

$$\begin{aligned} Q(\theta; \theta_t) :=& \mathbb{E}_{Z \sim p(z|X;\theta_t)}[l(\theta; X, Z)] \\ =& \int l(\theta; X, z) p(z|X; \theta_t) dz, \end{aligned} \tag{3.11}$$

where $p(z|X; \theta_t)$ is the conditional probability distribution of the unobserved variables Z conditioned to the observed variables X and the current fit of the parameters θ_t.

Then the **M-step** consists in finding the parameters θ that maximize the conditional expectation found in the E-step. This is,

$$\theta_{t+1} := \text{argmax}_\theta \, Q(\theta; \theta_t). \tag{3.12}$$

In our case, the latent variables are $\mathbf{r}$, $\mathbf{w}$ and $\mathbf{i}$. The observed variables are $\mathbf{y}$ (number of embryos implanted in each cycle), $\mathbf{x}$, $\mathbf{v}$ and $\mathbf{s}$. Then the conditional probability of the latent variables given the observed variables and the hyperparameters is

$$p(\mathbf{r}, \mathbf{w}, \mathbf{i}|\mathbf{y}, \mathbf{x}, \mathbf{v}, \mathbf{s}; \alpha, \beta, \theta) = \frac{p(\mathbf{y}|\mathbf{i}) p(\mathbf{r}, \mathbf{w}, \mathbf{i}|\mathbf{x}, \mathbf{v}, \mathbf{s}; \alpha, \beta, \theta)}{p(\mathbf{y}|\mathbf{x}, \mathbf{v}, \mathbf{s}; \alpha, \beta, \theta)} \tag{3.13}$$

where the denominator is given by Equation 3.10 and the numerator is equal to

$$p(\mathbf{y}|\mathbf{i}) p(\mathbf{i}|\mathbf{w}, \mathbf{r}, \mathbf{s}; \theta) p(\mathbf{w}|\mathbf{x}; \alpha) p(\mathbf{r}|\mathbf{v}; \beta) \tag{3.14}$$

With this equation we compute the weights corresponding to each cycle c, embryo e and configuration i that define the conditional distribution. Note that with Equations 3.8 and 3.9 we can decompose the probability into individual instances easily. For a given cycle c we have:

$$p(y_c|\mathbf{i}^c) p(\mathbf{i}^c|\mathbf{w}^c, r_c, \mathbf{s}^c; \theta) p(\mathbf{w}^c|\mathbf{x}^c; \alpha) p(r_c|v_c; \beta) \tag{3.15}$$

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. This is, we use Equation 3.15 and marginalize out the latent variables $\mathbf{w}$ and $\mathbf{i}$. Since the denominator does not depend on any latent variable we can use just the numerator and then normalize. We obtain the following expression:

$$q(r_c = r) \propto \Big(\sum_{\mathbf{i}^c \in \mathbb{I}_{\mathbf{s}^c, y_c}} \prod_e \sum_{w_e^c} p(i_e^c|w_e^c, r_c = r, s_e^c; \theta) p(w_e^c|x_e^c; \alpha) \Big) p(r_c = r|v_c; \beta). \tag{3.16}$$

Similarly, for each embryo e in the cycle we compute the weights corresponding to the two values of w_e^c, $w \in \{0, 1\}$. In this case we marginalize out r_c, $w_{e'}^c$ for any

$e' \neq e$ and $\mathbf{i}$.

$$q(w_e^c = w) \propto \sum_{r_c} \Bigg(\sum_{\mathbf{i}^c \in \mathbb{I}_{\mathbf{s}_:^c, y_c}} p(i_e^c | w, r_c, s_e^c; \theta) p(w | x_e^c; \alpha) \cdot \prod_{e' \neq e} \sum_{w_{e'}^c} p(i_{e'}^c | w_{e'}^c, r_c, s_{e'}^c; \theta) p(w_{e'}^c | x_{e'}^c; \alpha) \Bigg) p(r_c | v_c; \beta) \tag{3.17}$$

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

$$q(\mathbf{i}^c = \mathbf{i}) \propto \sum_{r_c} \Big(\prod_e \sum_{w_e^c} p(i_e | w_e^c, r_c, s_e^c; \theta) p(w_e^c | x_e^c; \alpha) \Big) p(r_c | v_c; \beta) \tag{3.18}$$

Note that if $\mathbf{i} \notin \mathbb{I}_{\mathbf{s}_:^c, y_c}$ then $p(y_c|\mathbf{i}) p(\mathbf{i} | \mathbf{w}^c, r_c, \mathbf{s}^c; \theta) = 0$ and the corresponding weight would be zero too.

Then the **M-step** consists in finding:

$$\text{argmax}_{\alpha, \beta, \theta} \, \mathbb{E}_{(\mathbf{w}, \mathbf{r}, \mathbf{i}) \sim q} \log p(\mathbf{r}, \mathbf{w}, \mathbf{i}, \mathbf{y} | \mathbf{x}, \mathbf{v}, \mathbf{s}; \alpha, \beta, \theta), \tag{3.19}$$

where q denotes the conditional probability described by the weights defined in Equations 3.16, 3.17 and 3.18. This conditional expectation has the following form:

$$\sum_c \sum_{\mathbf{i}^{c'} \in \mathbb{I}_{\mathbf{s}_:^c, y_c}} q(\mathbf{i}^{c'}) \Bigg[\sum_{r_{c'}} q(r_{c'}) \Big[\log p(r_{c'} | v_c; \beta) + \sum_e \sum_{w_{e'}^c} q(w_{e'}^c) \big[\log p(i_{e'}^c | w_{e'}^c, r_{c'}, s_e^c; \theta) + \log p(w_{e'}^c | x_e^c; \alpha) \big] \Big] \Bigg]. \tag{3.20}$$

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 3.16, 3.17 and 3.18, 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 re-train the probabilistic classifiers with the new weights obtained from the previous E-step. For θ_1, we maximize the conditional expectation of the log-likelihood given in Equation 3.19. The resulting maximum likelihood estimator is:

$$\hat{\theta}_1 = \frac{\sum_c \sum_{\mathbf{i}^{c'} \in \mathbb{I}_{\mathbf{s}_:^c, y_c}} \sum_e q(\mathbf{i}^{c'}) q(r_c = 1) q(w_e^c = 1) i_{e'}^c}{\sum_c \sum_{\mathbf{i}^{c'} \in \mathbb{I}_{\mathbf{s}_:^c, y_c}} \sum_e q(\mathbf{i}^{c'}) q(r_c = 1) q(w_e^c = 1)} \tag{3.21}$$

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 initializations to mitigate the local-maximum problem of EM algorithms.

Chapter 4

Experimental setup

The main goal of this project is to analyze the general probabilistic model proposed in Chapter 3 and compare the performance with different probabilistic classifiers. In particular, we study the effect in our model of the ASEBIR quality score (Ardoy and Calderon, 2008), and whether both our model and this score agree on the embryo selection. Moreover we use a set of alternative simpler models as a basic reference point to compare against.

In this chapter we explain all the details regarding the experimental setup used in our work. Special attention is given to the different probabilistic classifiers used in the models, the set of baseline models to compare and the evaluation procedure.

4.1 Probabilistic classifiers

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 methods, we use four different probabilistic classifiers:

- Logistic Regression (LR). Linear model that is based on maximum likelihood principles and whose predicted probabilities are modeled by a logistic function. Let (x, y) be a pair of features and label, then the predicted probability is:

$$p(y = 1|x) = \frac{1}{1 + e^{-(\theta_0 + \theta^T x)}}, \tag{4.1}$$

  where θ_0 and the vector θ are the parameters to estimate. The fitting process is performed maximizing the likelihood of the model for a training set. It is usual to apply regularization procedures, in particular we use an L2 loss with the strength parameter C, which results in the following optimization function:

$$\min_{\theta_0,\theta} \frac{1}{2}\theta^T\theta + C\sum_i \log\left(1 + e^{(-1)^{y_i}(\theta_0 + \theta^T x_i)}\right). \tag{4.2}$$

- Random Forest (RF200). Decisions Trees are a non-parametric supervised learning strategy that predict the value of a target variable learning simple decision rules on the input features. The basic idea is to partition the space into patches (using axis-orthogonal hyperplanes) and fit the model in those patches. This method is simple and easy to interpret but it is also prone to overfitting and very non-robust. Random Forest is an ensemble learning technique that creates many decision trees for different samples from the training set drawn with replacement (bootstrap samples). Furthermore, in each split only a random

subset of features is selected. Finally, the trees are aggregated and the predictions decided by majority voting. The two sources of randomness (bagging and feature selection) helps to decrease variance and overfitting and improve diversity.

- Extra-trees classifier (ETREES). Extremely Randomized trees add a further step of randomness with respect to Random Forests. As in RF, ETREES selects only a subset of candidate features. However, in the split for each candidate feature it does not only select the most discriminant threshold. Instead it draws random thresholds from the empirical distribution of the feature in training and the split with highest score is chosen.
- Gradient Boosting (GBOOST). Boosting techniques are also ensemble algorithms that use weak learners (such as shallow decision trees) to create stronger and more robust ones. Instead of optimizing all the learners at each step it uses a stage-wise approximation: the model up to current time is fixed and a new learner is added and optimized with respect to the data and the current model. Gradient Boosting allows for the optimization of any arbitrarily differentiable loss function. In particular we use a log loss as optimization function.

4.2 Baseline methods

To compare with our complete probabilistic model, we consider four different alternative methods. All of them use the probabilistic classifiers described in the previous section as direct predictors of whether an embryo will implant in the given cycle. The differences between them arise from the treatment of the unlabeled samples and the use or not of the cycles' features. These methods are designed in an incremental way, starting from a simple model where all unknown outcomes are fixed as negative to a more sophisticated one where the proportion of partially implanted cycles are used to infer information in the EM algorithm.

Method	Features	EM algorithm
Baseline_0	Embryo	No
Baseline_cycles	Embryo + Cycle	No
Naive EM	Embryo	Yes
EM w LP	Embryo	Yes (with label proportions)

4.2.1 Baseline_0 and Baseline_cycles

The first baseline method is the simplest both in terms of features and learning algorithm. Baseline_0 uses a probabilistic classifier to predict implantation just from the features of the embryos. This is in fact very similar to the embryo module of the general probabilistic model, which predicts the probability of the embryo being willing to implant: $p(w_e^c \mid x_e^c; \alpha)$. However, in this case the model predicts directly implantation since we use the implantation labels to train the model. Therefore we will compare it with the final outcome of the complete probabilistic model.

Similarly, the Baseline_cycles method uses a probabilistic classifier to predict implantation based on both the embryos and cycles features. Both sets of features are concatenated and fed to the model directly, hence no independence assumption is made in advance (unlike in our general probabilistic model). Nevertheless, note that some classifiers may assume independence by its own, such as Logistic Regression.

Both baseline models assume that all embryos with unknown label (those that were not transferred or whose cycle presented only partial implantation) belong to the negative class. Figure 4.1 shows the resulting partition of embryos regarding their label.

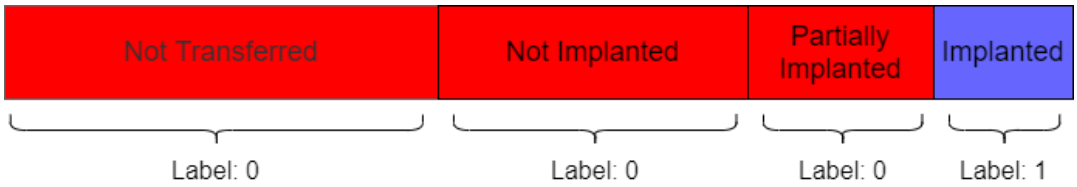

FIGURE 4.1: Labels of all the embryos in the dataset for the Baseline 0 and Baseline Cycles models.

With this assumption we obtain an even more severe class imbalance, with only a minute portion of embryos being labeled as positive (implanted). This could cause the model to predict a quite low general probability of implantation.

4.2.2 Naive EM

The second step in the incremental building of the methods is to incorporate an EM algorithm to account for the unlabeled or partially labeled embryos. In this model we use a simple EM algorithm where all embryos with unknown outcome are used in a semi-supervised procedure independently if they were transferred or not (Figure 4.2).

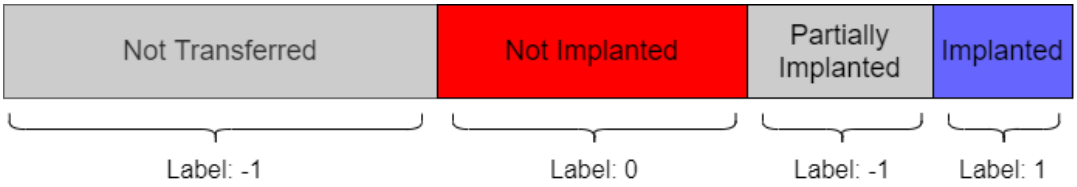

FIGURE 4.2: Labels of all the embryos in the dataset for the Naive EM model. Label -1 accounts for all the embryos with unknown outcome.

Let C_e be the random variable describing whether the embryo implants or not ($C_e = 1$ or $C_e = 0$). Then we define a pair of weights for each embryo $q(C_e = d)$ corresponding to the two possible values. These weights are computed as the conditional probability of obtaining $C_e = d$ given the model and features: $p(C_e = d \mid x_e; \alpha)$, where x_e are the features of the embryo and α is an hyperparameter. Then the learning algorithm is initialized with random weights for the unknown labels and the following two steps repeated iteratively:

Expectation: The unfixed weights are updated using the current fit of the model $(\hat{\alpha})$.

Maximization: The hyperparameter α is re-estimated retraining the classifier with the new weights obtained in the Expectation step.

The main drawback of this model is that all embryos with unknown outcome are treated equally. That is, embryos in partially implanted cycles are following the same learning process as non-transferred embryos. In other words, we are not using the whole information available.

4.2.3 EM with label proportions

In this last baseline method we add the information of label proportions in partially implanted cycles to the learning algorithm. Now the transferred embryos are considered in groups corresponding to each cycle. The number of implanted embryos is known for all cycles. For fully implanted cycles and failed cycles, this strategy does not add anything new since their associated embryos already knew their outcome. However, for embryos in partially implanted cycles it provides relevant information not considered before. Figure 4.3 shows how the embryo dataset is partitioned depending on their label.

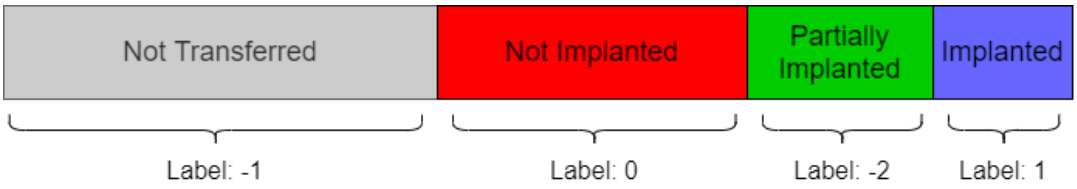

FIGURE 4.3: Labels of all the embryos in the dataset for the Naive EM model. Label -1 accounts for unlabeled embryos (Naive EM). Label -2 accounts for partially labeled embryos: only the label proportion is known.

For the non-transferred embryos the learning strategy is the same as in the Naive EM case. However, for the partially implanted embryos we compute the weights as the conditional probability of each class (given the features, the number of embryos implanted in the cycle and the hyperparameter). Let y_c be the number of embryos implanted on cycle c and i^c the vector describing the combination of embryos implanted ($i^c = [0,1,0,1]$ denotes that the second and fourth embryos implanted). Then we have:

$$p(i^c|x;\alpha) = \prod_{e \in S_c} p(C_e = i^c_e|x_e;\alpha) \tag{4.3}$$

where S_c is the set of embryos transferred in cycle c. Then the weights for an embryo $e \in S_c$ are computed as:

$$q(C_e = d) = \frac{\sum_{i^c \in \mathbb{I}_{y_c}} p(i^c|x;\alpha)\mathbb{1}[i^c_e = d]}{\sum_{i^c \in \mathbb{I}_{y_c}} p(i^c|x;\alpha)} \tag{4.4}$$

where $\mathbb{I}_{y_c}$ is the set of vectors compatible with the observed number of implanted embryos. That is, the combinations of implanted embryos in the cycle that result in y_c implantations.

4.3 Evaluation

Because of the weakly supervised nature of the problem (Hernández-González, Inza, and Lozano, 2016), 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 for learning 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:

$$p(i_e^c = 1|x_e^c, s_e^c, v_c; \alpha, \beta, \theta) = p(i_e^c = 1|w_e^c = 1, s_e^c, r_c = 1; \theta)p(w_e^c = 1|x_e^c; \alpha)p(r_c = 1|v_c; \beta). \tag{4.5}$$

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 $s_e^c = 0$, $p(i_e^c = 1|w_e^c, s_e^c = 0, r_c; \theta) = 0$. This is the reason why the evaluation is only performed with embryos which were transferred ($s_e^c = 1$). The other two terms in Eq. 4.5 are given by the probabilistic classifiers.

4.3.1 Performance metrics

Performance is assessed in terms of different metrics. Most of them are standard measures but require a particular interpretation of the results, since they are used in different context.

AUC-ROC

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 (Fawcett, 2006).

A Receiver Operating Characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR). It depicts the relative tradeoffs between TPR and FPR as the discrimination threshold is varied. A classifier with a very low threshold will predict as positive most of the instances, which results in a high TPR but also a high FPR. The ROC curve shows these different pairs of (FPR,TPR), known as operating points.

The advantage of ROC curves is that it gives a representation of the predictive performance of the classifiers independently of the threshold. From these curves we can obtain useful metrics such as the Area Under the Curve (AUC). This metric summarizes well the ROC behaviour since it accounts for the intuitive fact that for a given FPR we want the TPR to be as high as possible (higher curve). Moreover it has an important statistical property: the AUC score is equal to the probability that the classifier will rank a randomly chosen positive instance higher than a negative one. It is indeed a value between 0 and 1, and a higher score suggests a better classifier. Note that no realistic classifier should obtain a value less than 0.5 since this is the score obtained by a completely random classifier which is not able to distinguish positive and negative classes. This is represented by a diagonal line in the ROC curve (Figure 4.4).

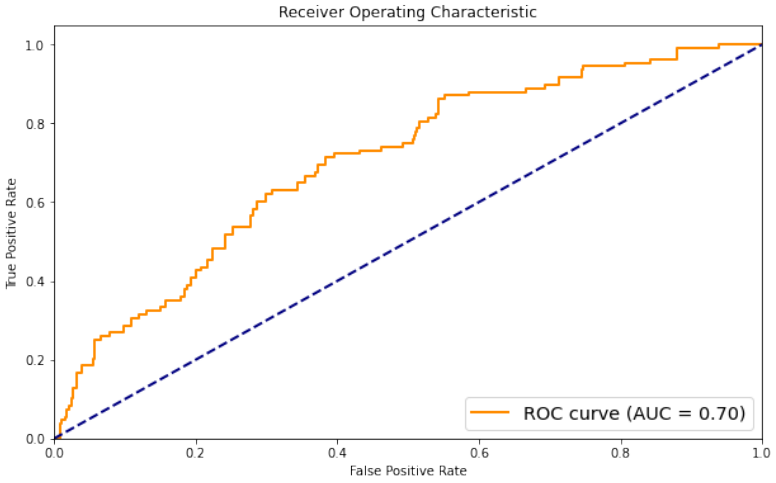

FIGURE 4.4: Example of a ROC curve.

LP-loss

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.

Let N_c be the number of transferred embryos in cycle c, and y_c the number of implanted ones. Then the LP-loss is

$$LP(\mathbf{Y};\alpha,\beta,\theta) = -\frac{1}{B}\sum_{c}\frac{|y_c - \hat{y}_c|}{N_c}, \tag{4.6}$$

where $\hat{y}_c$ is the number of implanted embryos predicted for cycle c and B the number of cycles.

Negative log-likelihood

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

$$\mathcal{L}(\mathbf{Y};\alpha,\beta,\theta) = -\frac{1}{B}\sum_{c=1}^{B}\sum_{j=0}^{N_c}\mathbb{1}[y_c = j]\log p(y_c), \tag{4.7}$$

where $p(y_c)$, the probability of cycle c having y_c implanted embryos, is,

$$p(y_c) = \sum_{i^c \in \mathbb{I}_{y_c}}\prod_{e}[i_e^c p(i_e^c = 1) + (1 - i_e^c)p(i_e^c = 0)] \tag{4.8}$$

where $p(i_c)$ is given by Eq. 4.5 and $\mathbb{I}_{y_c}$ consists of the possible joint assignment of value (vector) to all the $\{i_e^c\}_{e\in E_c}$, as explained in the context of Eq. 3.10.

Chapter 5

Results

In this chapter we show the results obtained for the different experiments performed: with our complete probabilistic model and with the baseline models. The section is divided in two main parts. In the first one (Section 5.1) we explore the results of the main probabilistic model and give special attention to its relation with the ASEBIR score. In particular we show the effect of including this score as a feature in the model. Then the behaviour of the model is further analyzed with the help of the predicted probability densities; we study the separation of the output depending on three factors: implantation of the pair embryo-cycle, transfer of the embryo and ASEBIR score of the embryo. In the second part (Section 5.2) we compare the results of the probabilistic graphical model with the ones obtained for the baseline models. Not only the performance of all models is addressed but also the aforementioned predicted probability densities.

5.1 Probabilistic graphical model and effect of ASEBIR

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 5.1 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 (Ardoy and Calderon, 2008).

In terms of performance, GBOOST and RF seem to be the best 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 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 perform relatively well in a pure classification task (implantation or not), but it fails on estimating the probability of more uncertain cases.

Table 5.2 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.

TABLE 5.1: Metrics and control measures obtained using 5-fold cross validation

Model	Classifier	AUC	lp_loss	loglikelihood
Full Model	ETREES	0.64 ± 0.07	$\mathbf{0.54 \pm 0.05}$	1.45 ± 1.59
	GBOOST	$\mathbf{0.71 \pm 0.04}$	0.72 ± 0.03	0.45 ± 0.05
	LR	0.63 ± 0.08	0.60 ± 0.05	0.51 ± 0.10
	RF	$\mathbf{0.71 \pm 0.05}$	0.80 ± 0.05	$\mathbf{0.42 \pm 0.07}$
Full Model (Hidden quality)	ETREES	0.64 ± 0.05	$\mathbf{0.54 \pm 0.05}$	1.27 ± 1.57
	GBOOST	$\mathbf{0.73 \pm 0.07}$	0.73 ± 0.07	0.43 ± 0.06
	LR	0.62 ± 0.08	0.64 ± 0.07	0.52 ± 0.10
	RF	0.71 ± 0.05	0.80 ± 0.05	$\mathbf{0.42 \pm 0.07}$

TABLE 5.2: 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
	RF	0.48 ± 0.01		RF	0.48 ± 0.01

For the GBOOST, LR and RF 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 5.1).

To fully grasp the behaviour of the models, we also analyze the different predicted probability densities output by them. Figure 5.1 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 5.1, 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 (see Figure 2.3), 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 5.1, 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.

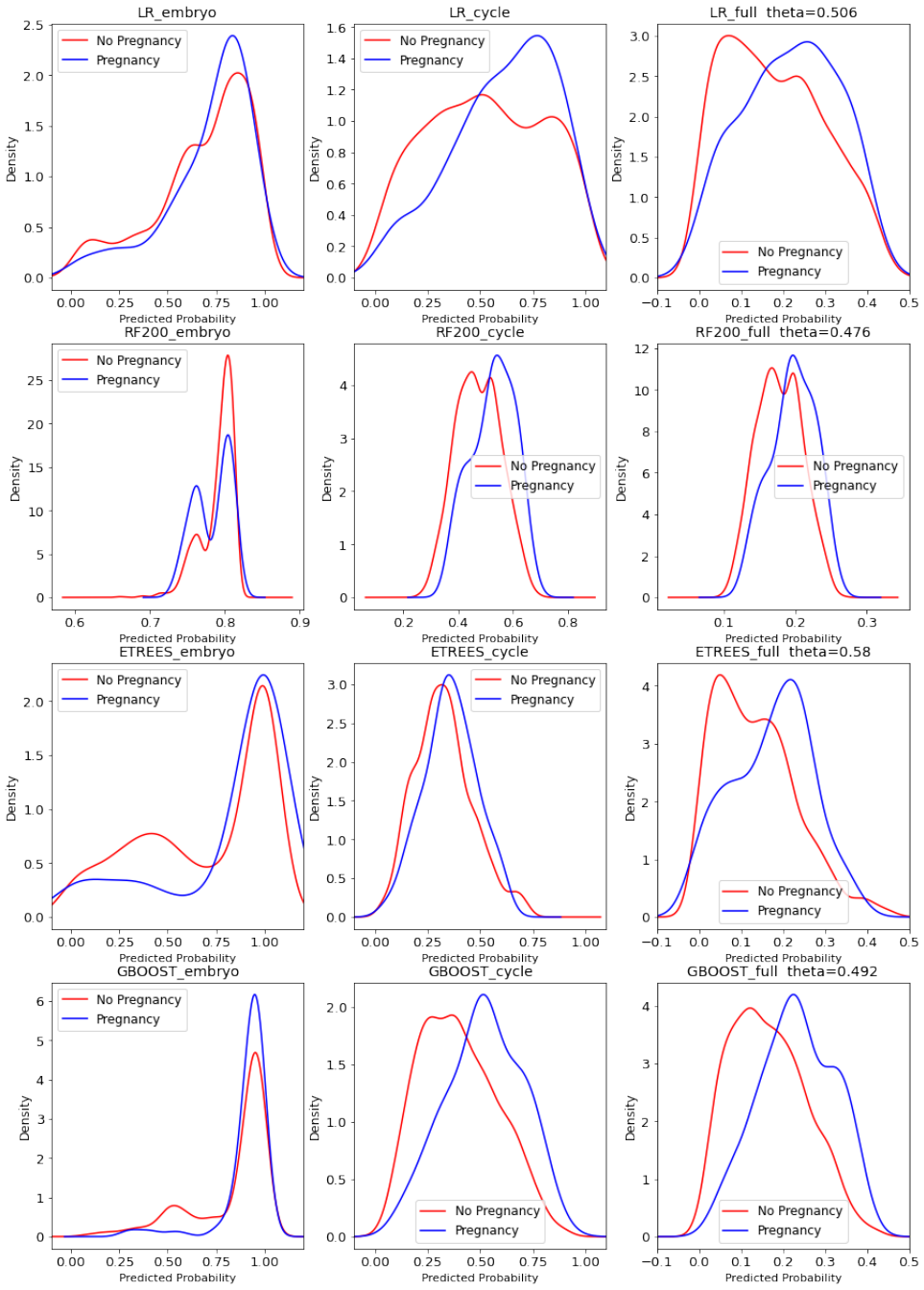

FIGURE 5.1: Density of the predicted probabilities for an embryo to be willing to implant, for a cycle to be willing to let embryos implant and for the pair embryo-cycle to actually induce pregnancy. The figure shows the different probability densities depending on the true outcome of each embryo-cycle pair (induce pregnancy or not). Each row corresponds to a different probabilistic classifier (ETREES, GBOOST, LR and RF).

Figure 5.2 shows, in a similar way, the probability densities for each embryo-cycle pair separating on whether the embryos were transferred. The left column indicates that transferred embryos are predicted to have a higher probability of being willing to implant. This means that the model agrees with the selection for transfer. The middle column shows that in pairs where the embryo is not transferred

the probability of the cycle being viable is slightly larger. It is important to note that this probability is independent of the actual individual embryo; the probability of a cycle being viable is the same for all embryos associated to it (transferred or not). However, there is a bias inducing factor to consider. The number of embryos created in a cycle is thought to be correlated with the fertility of the uterus: viable cycles produce more embryos. The number of transferred embryos is similar for all cycles but the number of not transferred ones is larger for those cycles with many embryos. This means that there are more not transferred embryos corresponding to fertile uterus, which results in the aforementioned bias. This also causes the right column (probability of implantation) to be biased.

Likewise, Figure 5.3 displays 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 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. However, for the RF classifier the quality of the embryo does not seem to be a decisive factor. We can see in this case that the probability of an embryo being good has a strange two-peak relation with quality but it does not actually translate into a relevant separation for the final prediction. For the other 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*.

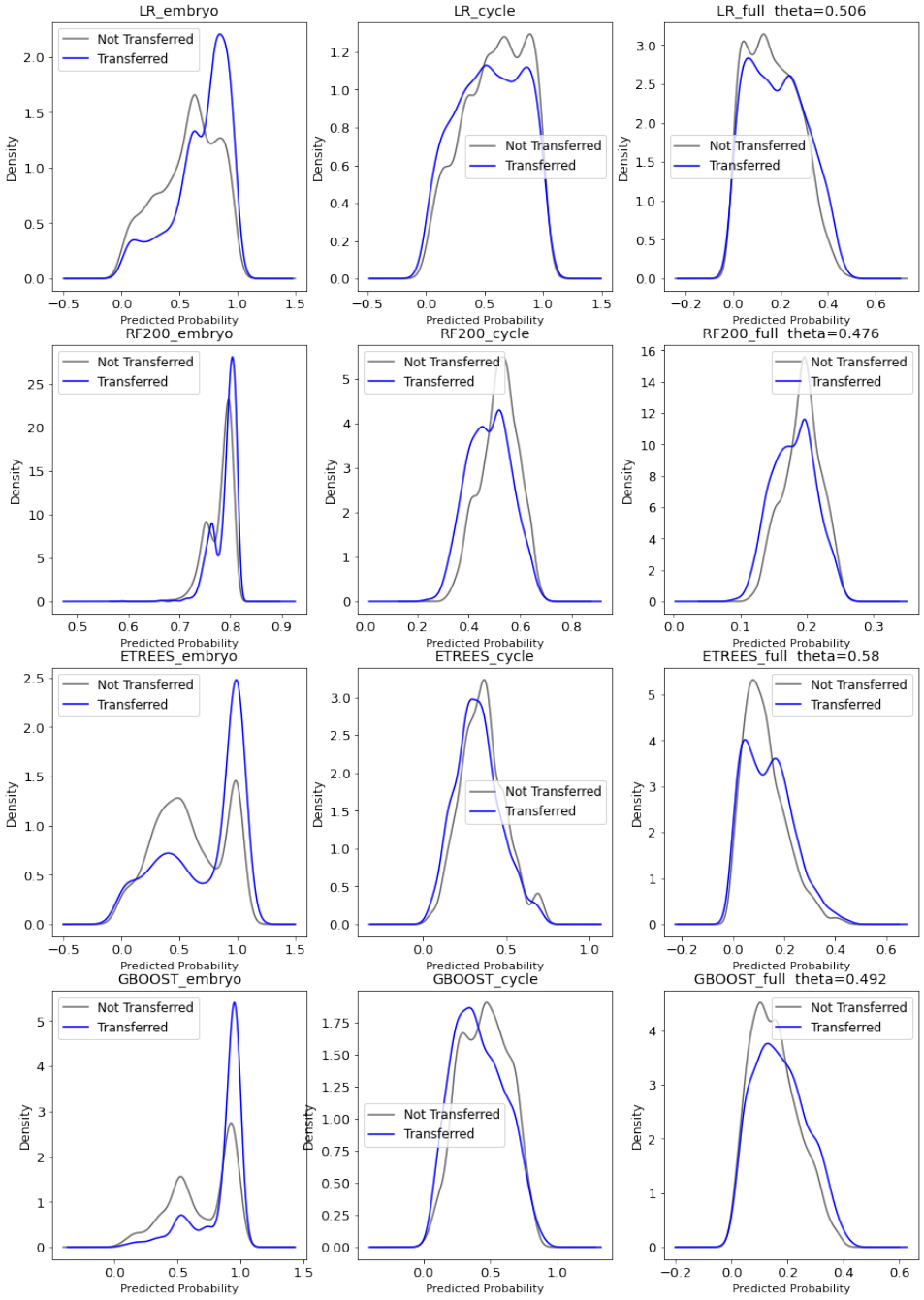

FIGURE 5.2: Density of the predicted probabilities for an embryo to be willing to implant, for a cycle to be willing to let embryos implant and for the pair embryo-cycle to actually implant. The figure shows the different probability densities depending on whether the embryo was transferred or not. Each row corresponds to a different probabilistic classifier (ETREES, GBOOST, LR and RF).

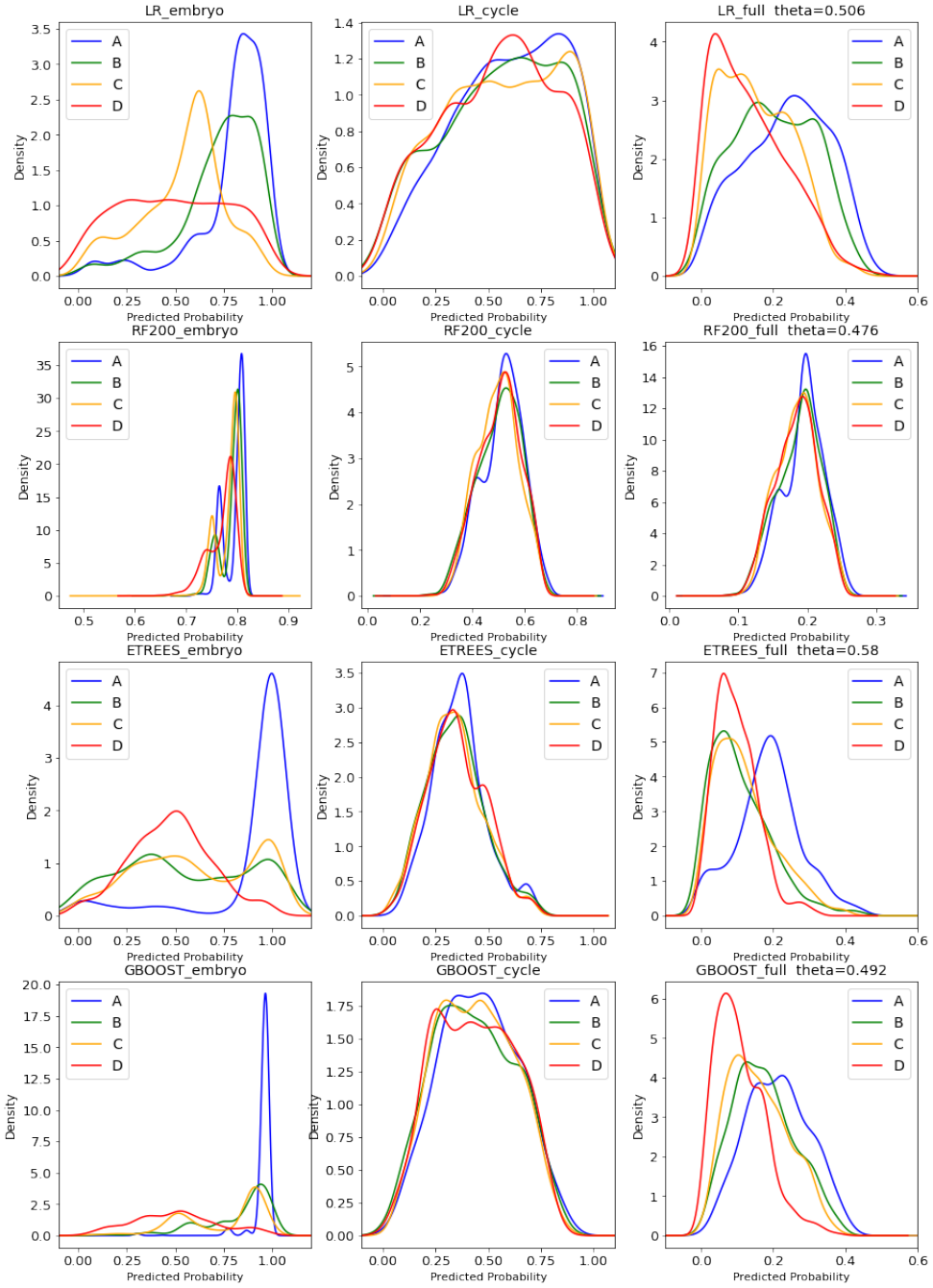

FIGURE 5.3: Density of the predicted probabilities for an embryo to be willing to implant, for a cycle to be willing to let embryos implant and for the pair embryo-cycle to actually implant. The figure shows the different probability densities depending on the ASEBIR quality score given to the embryo (A, B, C or D). Each row corresponds to a different probabilistic classifier (ETREES, GBOOST, LR and RF).

5.2 Comparison with baseline methods

In this part we compare the performance of our full probabilistic model with the baseline methods described in Section 4.2. Table 5.3 shows the metrics computed for all methods with the different probabilistic classifiers. All the models were trained without the ASEBIR score feature.

TABLE 5.3: Metrics obtained using 5-fold cross validation for the probabilistic graphical model and the baseline models.

Method	Classifier	AUC	lp_loss	loglikelihood
Baseline_0	ETREES	0.64 ± 0.07	$\mathbf{0.20 \pm 0.05}$	∞
	GBOOST	0.62 ± 0.05	0.21 ± 0.05	0.67 ± 0.26
	LR	0.58 ± 0.06	0.21 ± 0.05	0.63 ± 0.18
	RF	0.61 ± 0.06	$\mathbf{0.20 \pm 0.05}$	0.62 ± 0.18
Baseline_cycles	ETREES	0.62 ± 0.05	$\mathbf{0.20 \pm 0.05}$	∞
	GBOOST	0.72 ± 0.05	$\mathbf{0.20 \pm 0.05}$	0.64 ± 0.12
	LR	0.63 ± 0.07	$\mathbf{0.20 \pm 0.05}$	0.70 ± 0.25
	RF	$\mathbf{0.74 \pm 0.06}$	$\mathbf{0.20 \pm 0.05}$	0.58 ± 0.15
Naive EM	ETREES	0.50 ± 0.08	0.27 ± 0.04	∞
	GBOOST	0.61 ± 0.08	0.21 ± 0.05	0.51 ± 0.12
	LR	0.56 ± 0.06	$\mathbf{0.20 \pm 0.05}$	0.51 ± 0.11
	RF	0.55 ± 0.07	$\mathbf{0.20 \pm 0.05}$	0.46 ± 0.11
EM w LP	ETREES	0.50 ± 0.09	0.28 ± 0.04	∞
	GBOOST	0.60 ± 0.08	0.21 ± 0.05	0.44 ± 0.05
	LR	0.56 ± 0.06	$\mathbf{0.20 \pm 0.05}$	0.47 ± 0.05
	RF	0.58 ± 0.08	$\mathbf{0.20 \pm 0.05}$	$\mathbf{0.42 \pm 0.06}$
Full PGM	ETREES	0.64 ± 0.05	0.54 ± 0.05	1.27 ± 1.57
	GBOOST	0.73 ± 0.07	0.73 ± 0.07	0.43 ± 0.06
	LR	0.62 ± 0.08	0.64 ± 0.07	0.52 ± 0.10
	RF	0.71 ± 0.05	0.80 ± 0.05	$\mathbf{0.42 \pm 0.07}$

In terms of AUC-ROC score, both the Baseline_cycles and the full PGM obtain the best performance. As mentioned in the last section, the AUC-ROC is calculated using only embryos whose outcome is known. Therefore it gives a measure of how well the models perform in situations where either all embryos implanted or none did. In that sense, it is reasonable to think that in those cases the cycle is the critical factor for the outcome (it is in general but more so in these cases). Indeed, the Baseline_cycles and Full PGM are the only models that use the cycles' features.

For the LP-loss, the best results are the ones obtained by the baseline models. However, the numbers are a bit misleading. If a model predicts all instances as failures it would already get a 0.20 LP-loss. This is the case for most of the baseline models, which predict quite low probabilities of implantation for all embryos. If the classification threshold on the probabilistic classifier is not taken properly, all the predictions are negative, resulting in an apparently good LP-loss. Therefore this metric does not provide much relevant information about the predictive power of the models.

Instead, the negative log-likelihood uses directly the output probability of implantation, which is independent of any threshold. Moreover, it takes into account the confidence of the model in the predictions. In this metric, we do not only use the embryos with known outcome: the partially implanted cycles are also taken into account. This is the reason why the models with an EM learning algorithm perform

better for this metric. Baseline_0 and Baseline_cycles completely disregard the information provided from partially implanted cycles (they are directly assigned to the negative class). Meanwhile, the EM models use this partially labeled cycles to learn. In the Naive EM case they are only used in a semi-supervised strategy, as if no information was actually available from the number of embryos implanted. However, in the EM with Label Proportions model the partial labels of the cycles is used in a weakly-supervised context. In fact, this extra information provides better results, comparable to the ones obtained by the full PGM.

5.2.1 Probability distributions

As in the last section, we show the output predicted probability densities separated depending of three factors: (i) the true label (implanted or not), (ii) whether the embryo was transferred and (iii) the embryo quality given by the ASEBIR score. It is important to note that most of the baseline methods use only the embryo features. Although it may seem that they are the equivalent to the embryo module of our Full PGM, these models are actually trying to predict implantation and not just viability of the individual embryo. The difference resides on the target variables. In the baseline methods we use directly the labels to train the model, which define embryo implantation. However, in the PGM we used the latent variable w_e which accounts for whether the embryos is willing to implant.

Figure 5.4 shows the probability densities for the Baseline_0 model. Here we see that for all the classifiers the predictions are extremely pessimistic. In fact, none of them gives a positive predicted probability even close to 0.5. This means that the predicted labels are all negative, hence the high result in the LP-loss (see Table 5.3). Nonetheless we see some signs of learning, even if the values are low. For instance, true positives are predicted to implant with slightly more probability in most of the classifiers, especially in the ETREES case. We can see even clearer that embryos with quality A are predicted with higher probability than the rest of embryos. In fact, quality D embryos are almost always predicted with probability 0. The ETREES classifier seems to predict the A quality embryos with higher probability than the rest of classifiers. However it does not show significant increases for middle quality embryos with respect to the poorest quality ones. Instead, LR and RF classifier do show some improvements for B and C quality embryos.

Similarly, Figure 5.5 shows the probabilities for the Baseline_cycles model. In this case the densities are smoother, but present a similar behaviour. With this model, the density of true negative embryos is clearly peaked close to 0 for all classifiers. Even though the mode of the true positive class is also close to the same value, we see many more of those embryos that are predicted with higher probability. This is enough to obtain a significantly higher AUC score than the Baseline_0 model, as seen in Table 5.3. For the separation by transfer, we observe a similar behaviour, with the non-transferred embryos being concentrated close to 0 and the transferred ones more spread to higher probabilities, although not as much as in the left column. The last column also shows some bias in favor of quality A embryos. Here we see more noticeable discrepancies between classifiers. For instance, RF does not present as much dispersion in the probabilities as the other classifiers. In general, the differences between embryo qualities are not as clear as in the Baseline_0 method. This may be due to the addition of the cycles features, which may have diluted the effect of the embryo morphological characteristics.

With the implementation of the EM strategy we start to see some relevant differences, even in the Naive version. Figure 5.6 shows the probabilities for the Naive

EM method. We can see that the predicted probabilities are a bit higher than in the previous cases. It is specially noteworthy the densities depending on the ASEBIR quality (right column). With this model we can see a more evident difference in the treatment of different quality embryos (recall that this score is not directly given to the classifiers). It is clear that the model predicts with higher probability the quality A embryos, while quality D embryos are predicted with less probability. This is specially visible for the LR and GBOOST classifiers. RF presents a similar behaviour but with less separation between qualities. However, the ETREES classifier has a strange behaviour regarding D quality embryos, which are predicted with higher probability than all the other ones. Looking at the middle column, we see that many non-transferred embryos are predicted positive with high probability. Moreover, recall from Chapter 2 that almost all of the quality D embryos were not transferred.

Finally, Figure 5.7 shows the predicted probability densities for the EM with label proportions model. The results seem to be more refined specially considering the separation by embryo quality. For the GBOOST and LR classifiers, we observe a clear difference between the highest quality A, the middle qualities B and C (which seem to be somewhat mixed) and the poorest quality D. However, there is still a strange behaviour in the ETREES classifier, giving considerable higher probability to implant to D quality embryos. Similarly, non-transferred embryos are predicted with higher probability than transferred ones. The rest of classifiers appear to separate well the embryos by transfer, especially GBOOST. This is also the case in the separation by the true label (left column), where GBOOST shows the best separation.

In general, we see that the implementation of the EM strategy helps to separate embryos depending on the ASEBIR quality, although not as much as the embryo module from our full probabilistic model. The baseline methods with this learning algorithm obtain considerably higher predicted probabilities and better negative log-likelihood (see Figure 5.3). Of course, the simplification of assuming all unknown outcomes as negative resulted in quite low probabilities. Nonetheless, the AUC score was not too affected by this. In fact, with the inclusion of the cycles features Baseline_cycles obtained similar scores to our full probabilistic model. Regarding the different classifiers, GBOOST and RF obtained the best performance metrics. However, in terms of separation by probability distributions, GBOOST presented much promising results, with significant differences for all the separation factors, especially for the EM with LP method.

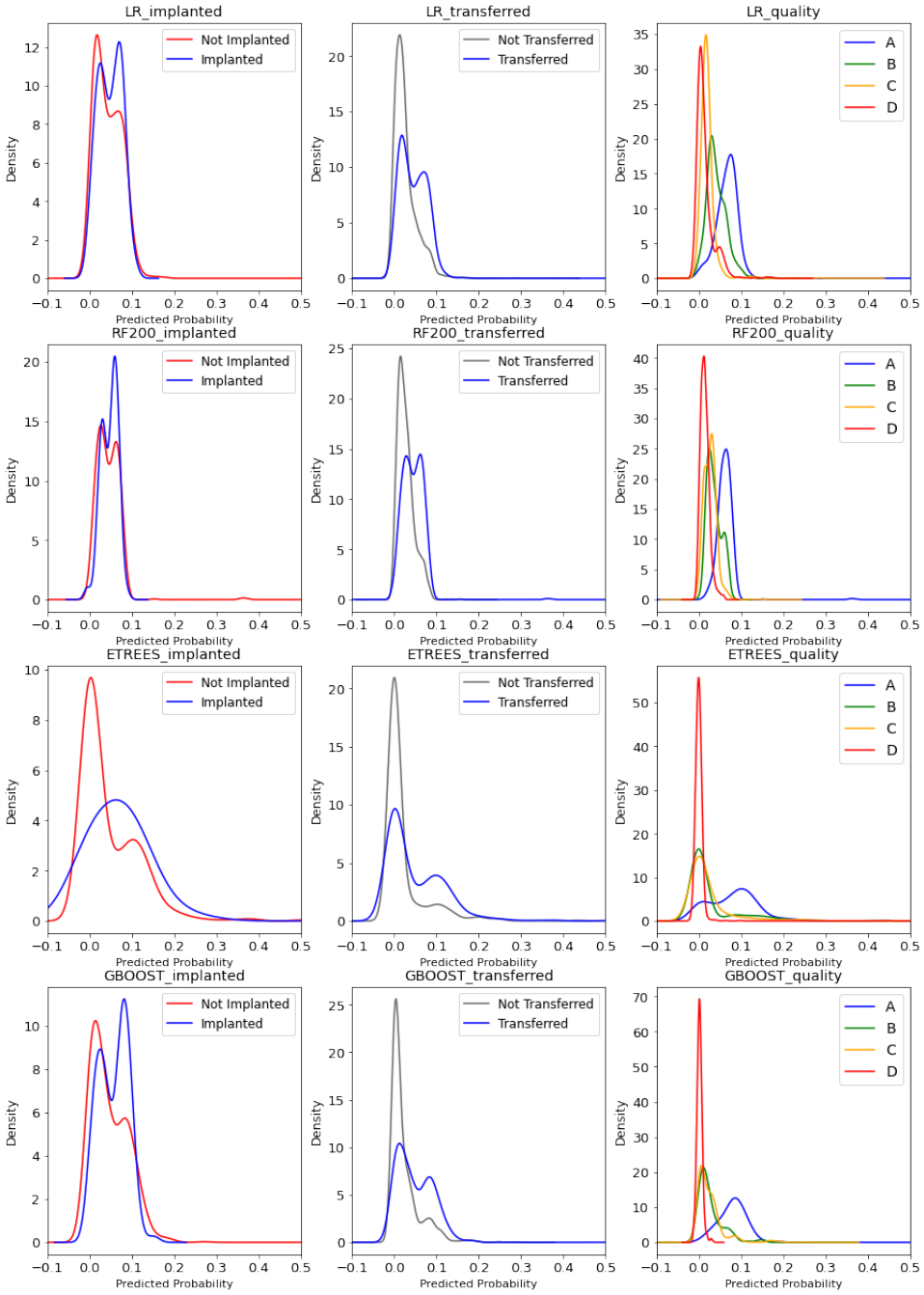

FIGURE 5.4: Density of the predicted probabilities for the Baseline_0 model. They are separated depending (i) on the true label (left column), (ii) on whether the embryo was transferred (middle column) (iii) and on the the embryo ASEBIR quality (right column).

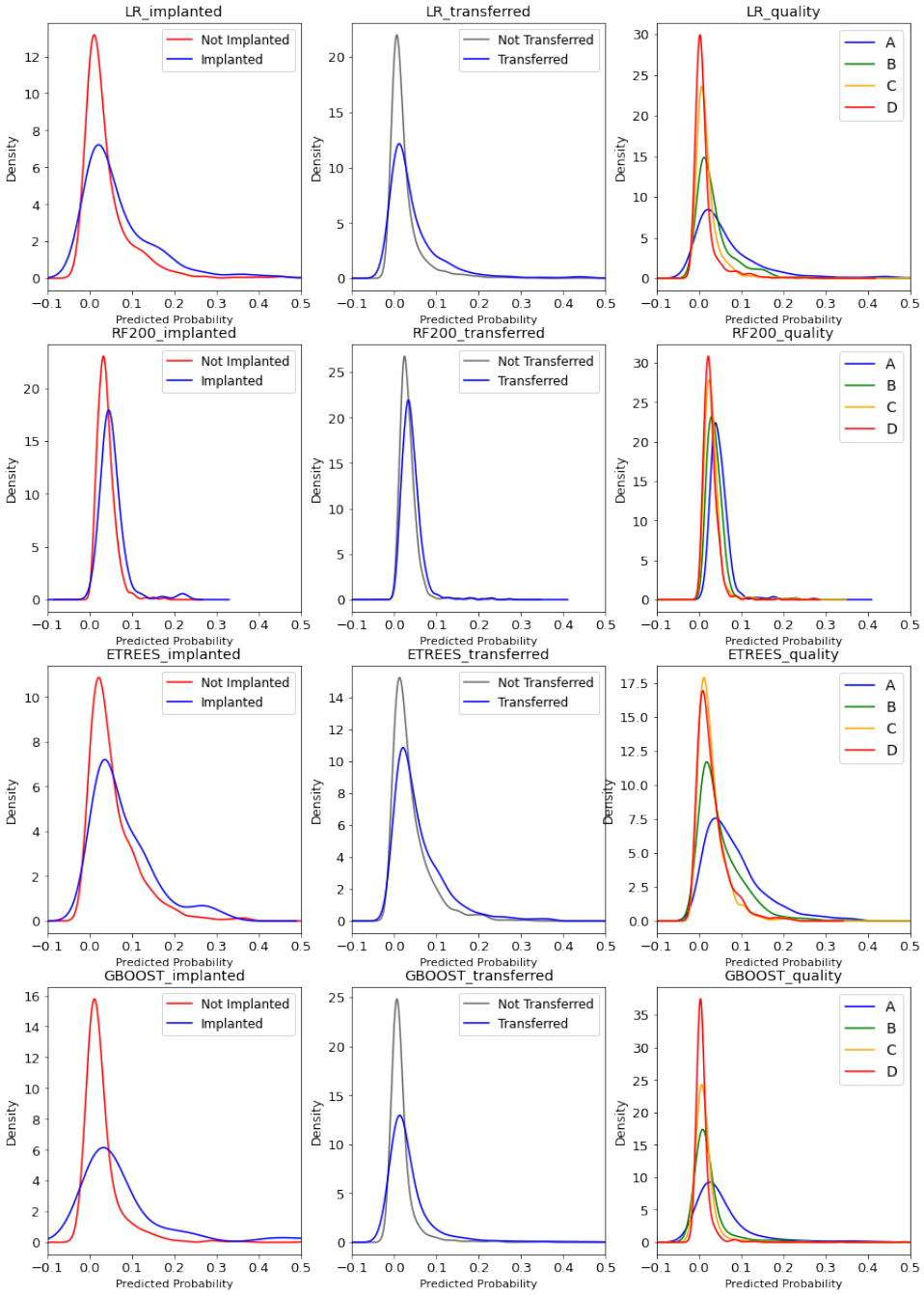

FIGURE 5.5: Density of the predicted probabilities for the Baseline_cycles model. They are separated depending (i) on the true label (left column), (ii) on whether the embryo was transferred (middle column) (iii) and on the the embryo ASEBIR quality (right column).

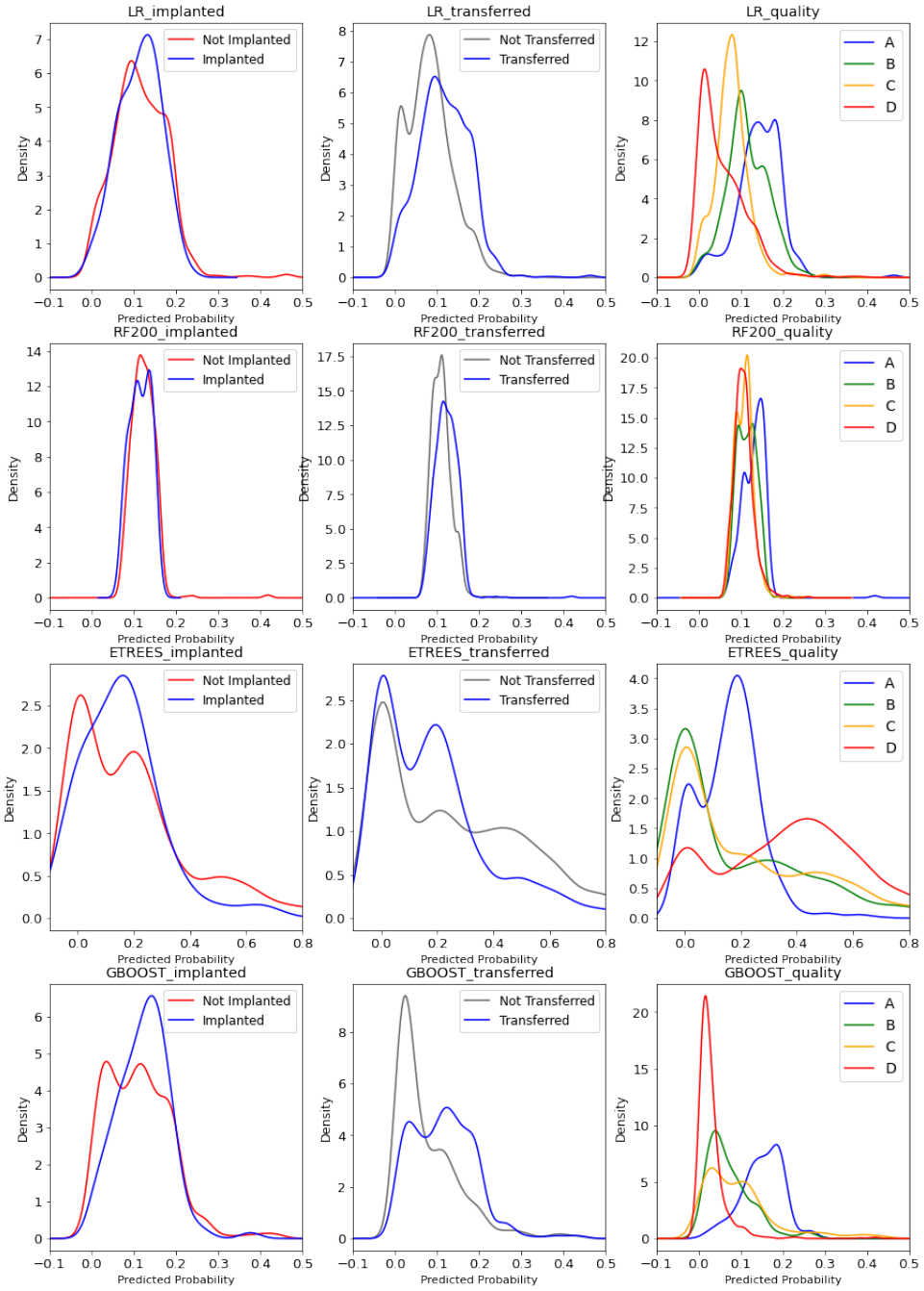

FIGURE 5.6: Density of the predicted probabilities for the Naive EM model. They are separated depending (i) on the true label (left column), (ii) on whether the embryo was transferred (middle column) (iii) and on the the embryo ASEBIR quality (right column).

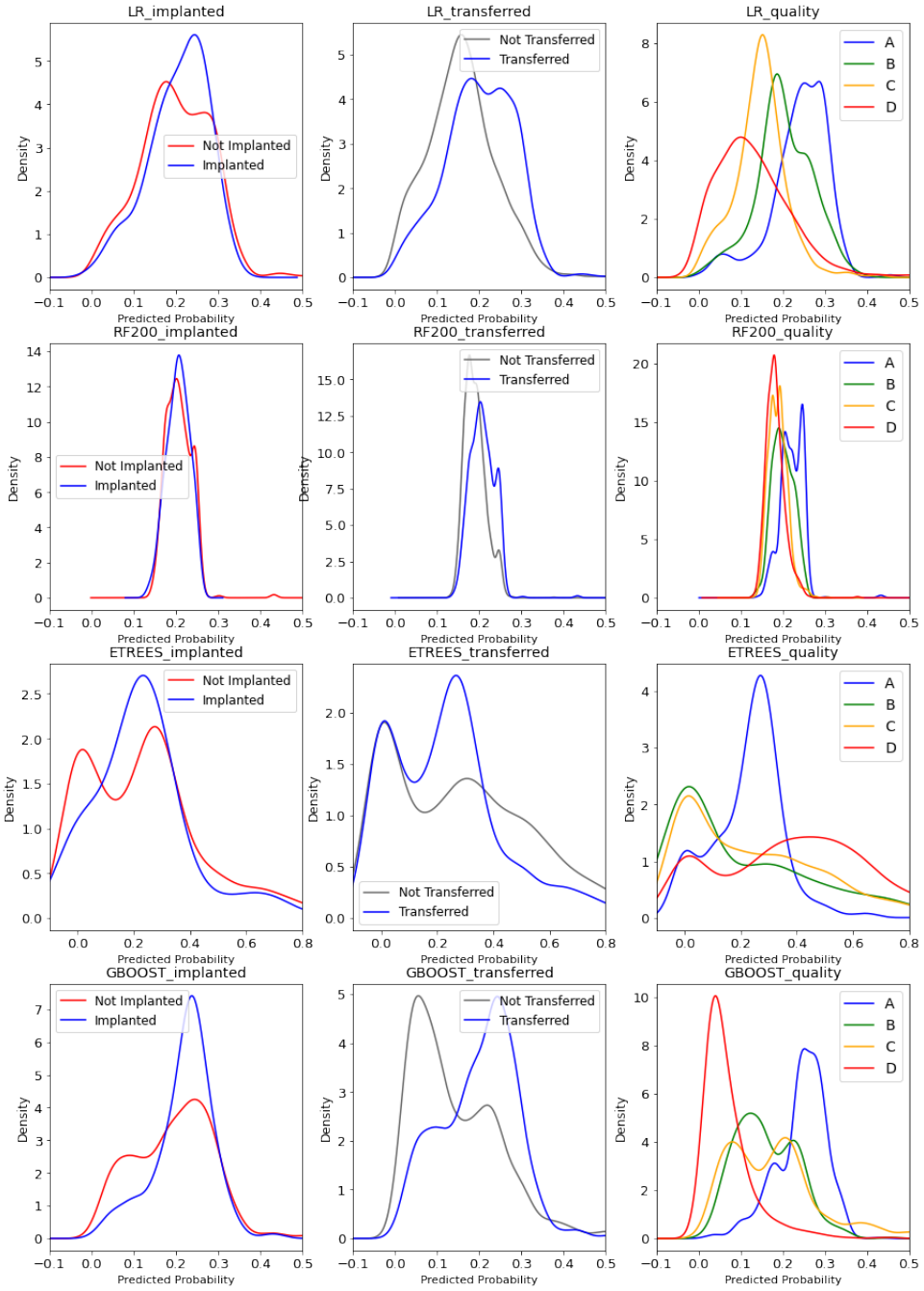

FIGURE 5.7: Density of the predicted probabilities for the EM with LP model. They are separated depending (i) on the true label (left column), (ii) on whether the embryo was transferred (middle column) (iii) and on the the embryo ASEBIR quality (right column).

Chapter 6

Conclusions

In this work, we address the problem of 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 implantation. The performance of the model is tested using different classifiers which evaluate the goodness of the embryos and cycles. Gradient Boosting and Random Forrest classifiers 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.

The performance of the model was further validated by an extended experimental setup where alternative, simpler baseline models were used to test different properties of the model. For instance we saw that the inclusion of the characteristic of the cycle is a key factor to predict implantation, especially in definitive cases (where either all embryos in a cycle implanted or none implanted). On another note we showed the benefits of implementing an EM strategy for the learning process. This was helpful to predict the actual number of implanted embryos in every cycle, even if the identity of the individual embryos is unknown. This resulted in a good separation of embryos by quality based on the predicted probability to implant.

Our complete probabilistic model brings together all these good properties while also providing some helpful features. For instance, it allows to assess the quality of individual embryos extracting the embryo module directly. We saw that this module predicted clearly that embryos with good quality score would implant with more probability than worse quality embryos. Moreover we obtained an estimation of the uncertainty originated from unknown, external factors. Most of the classifiers predicted than even when the embryo is deemed as willing to implant and the cycle as viable, there is only around 50% probability of actually inducing pregnancy.

6.1 Further Research

There are different research lines open after this exploration of the behaviour of our model in relation to the ASEBIR protocol. The experimental setup could be further increased considering other baseline models which, for instance, combine the

EM algorithm also with the cycles characteristics. 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.). Moreover, the treatment of the classifiers could be improved to obtain better predictive power. In this work we only used standard versions of them, without any tuning, since the objective was to compare different models with a simple setup.

Bibliography

Ardoy, M. and G. Calderon (2008). "Clinical Embryology Papers: ASEBIR criteria for the morphological evaluation of human oocytes, early embryos and blastocysts". In: *Asociación para el Estudio de la Biología de la Reproducción (ASEBIR)*.

Corani, G et al. (Nov. 2013). "A Bayesian network model for predicting pregnancy after in vitro fertilization". In: *Computers in biology and medicine* 43, pp. 1783–92.

Coughlan, C et al. (2015). "Recurrent implantation failure: definition and management". In: *Reproductive BioMedicine Online* 28.1, pp. 14–38.

Cuevas-Sáiz, I et al. (Feb. 2018). "The Embryology Interest Group: updating ASEBIR's morphological scoring system for early embryos, morulae and blastocysts". In: *Medicina Reproductiva y Embriología Clínica* 5.

Debón, A. et al. (2013). "Mathematical methodology to obtain and compare different embryo scores". In: *Mathematical and Computer Modelling* 57.5, pp. 1380–1394. ISSN: 0895-7177.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38.

Engmann, L et al. (Dec. 2001). "Trends in the incidence of births and multiple births and the factors that determine the probability of multiple birth after IVF treatment". In: *Human Reproduction* 16, pp. 2598–605.

Fawcett, T (2006). "Introduction to ROC analysis". In: *Pattern Recogn. Lett.* 27, pp. 861–874.

Fernandez, Eleonora et al. (Oct. 2020). "Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data". In: *Journal of Assisted Reproduction and Genetics* 00, p. 1.

Guérif, F. et al. (2007). "Limited value of morphological assessment at days 1 and 2 to predict blastocyst development potential: a prospective study based on 4042 embryos." In: *Human reproduction* 22 7, pp. 1973–81.

Hernández-González, J et al. (2018). "Fitting the data from embryo implantation prediction: Learning from label proportions". In: *Statistical Methods in Medical Research* 27, pp. 1056 –1066.

Hernández-González, J, I Inza, and J A Lozano (2016). "Weak supervision and other non-standard classification problems: A taxonomy". In: *Pattern Recogn. Lett.* 69, pp. 49–55.

Kragh, M et al. (2019). "Automatic grading of human blastocysts from time-lapse imaging". In: *Comput. Biol. Med.* 115, p. 103494.

Manna, C et al. (May 2004). "Experimental results on the recognition of embryos in human assisted reproduction". In: *Reproductive biomedicine online* 8, pp. 460–9.

Morales, Dinora A. et al. (June 2008). "Bayesian classification for the selection of in-vitro human embryos using morphological and clinical data". In: *Computer methods and programs in biomedicine* 90, pp. 104–16.

Patrizi, Giacomo et al. (July 2004). "Pattern recognition methods in human-assisted reproduction". In: *International Transactions in Operational Research* 11, pp. 365 – 379.

Racowsky, C. et al. (2009). "Is there an advantage in scoring early embryos on more than one day?" In: *Human reproduction* 24 9, pp. 2104–13.

Report, ESHRE Campus Course (Apr. 2001). "Prevention of twin pregnancies after IVF/ICSI by single embryo transfer". In: *Human Reproduction* 16.4, pp. 790–800.

Roberts, Stephen (Jan. 2007). "Models for assisted conception data with embryo-specific covariates". In: *Statistics in medicine* 26, pp. 156–70.

Siristatidis, Charalampos et al. (Mar. 2011). "Artificial intelligence in IVF: A need". In: *Systems biology in reproductive medicine* 57, pp. 179–85.

Speirs, Andrew L. et al. (1983). "Analysis of the benefits and risks of multiple embryo transfer". In: *Fertility and Sterility* 39.4, pp. 468–471. ISSN: 0015-0282.

Uyar, Asli, Ayse Bener, and H Nadir Ciray (May 2014). "Predictive Modeling of Implantation Outcome in an In Vitro Fertilization Setting: An Application of Machine Learning Methods". In: *Medical decision making : an international journal of the Society for Medical Decision Making* 35.

Valls Murcia, O (2021). "A comprehensive probabilistic model for the embryo selection problem". MA thesis. Technical University of Catalonia.

Zhou, Haibo and Clarice R. Weinberg (1998). "Evaluating effects of exposures on embryo viability and uterine receptivity in in vitro fertilization". In: *Statistics in Medicine* 17.14, pp. 1601–1612.