ASSESSING AND ENHANCING MACHINE LEARNING METHODS IN IVF PROCESS: PREDICTIVE MODELING OF IMPLANTATION AND BLASTOCYST DEVELOPMENT

by

Aslı Uyar Özkaya

B.S., in Control and Computer Engineering, İstanbul Technical University, 2003

M.S., in Computer Engineering, Boğaziçi University, 2006

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2011

Dedicated to our sons Eren, Bora, Can

and to all IVF babies

ACKNOWLEDGEMENTS

I owe my deepest gratitude to my advisor Assist. Prof. Ayşe Bener for her invaluable guidance and contributions to my research. She always made the right decision in critical milestones and motivated me to conduct target oriented research. I learned much from her not only academically but in every aspect of life.

I am grateful to Assoc. Prof. H. Nadir Çıray who provided all the necessary material and substantial knowledge in the medical side of this work. I would like to thank him for the insightful conversations and helpful suggestions during the process.

I would like to thank Prof. Lale Akarun for her valuable comments and feedbacks that improved the work; Prof. H. Levent Akın and Prof. Oğuz Tosun for kindly accepting to be a member of my defense jury.

This work would not have been completed without financial support of Bahceci IVF Clinic and Boğaziçi University Research Fund (BAP 09A104D). I thank to Osman Arslan for his effort in collection of IVF data. I thank to members of Softlab; Ayşe Tosun Mısırlı, Bora Çağlayan and Gül Çalıklı for our scientific discussions and for all the good time we had and Burak Turhan for all the references he provided that facilitated my research. I also thank to Levent Özgür, Yasemin Şengül, Deniz Demirhan Barı, Itır Karaç and Çetin Meriçli for their sincere friendship and moral support when I needed most.

I am also grateful to my sisters who took care of my son when I was busy. I am forever indebted to my husband Fatih, who supported me to the end with incredible patience and endless love. Lastly and most importantly, I wish to thank my son Eren, for his innocent motivation and of course for his assistance in printing and stapling my documents.

ABSTRACT

ASSESSING AND ENHANCING MACHINE LEARNING METHODS IN IVF PROCESS: PREDICTIVE MODELING OF IMPLANTATION AND BLASTOCYST DEVELOPMENT

In this thesis, we address the decision-making problems in in vitro fertilization treatment from the machine learning perspective aiming to increase the clinical success rates. Initially, we present a comprehensive and comparative analysis of the classification techniques in embryo-based implantation prediction. In parallel, we evaluate the predictor effects of input features in order to eliminate the redundant variables and decide the optimum feature subset leading to the highest prediction performance. In contrast to the limited relevant literature, our preliminary experiments demonstrate the potential of machine learning classifiers as an automated decision support tool in critical decisions affecting the success of the treatment. Later, we focus on improving the classification performance either by algorithmic enhancements or by improving the information content of the data. First, we handle the problem of imbalanced class distribution and show that decision threshold optimization and re-sampling the training data produce similar results. Second, we propose a frequency based encoding technique to efficiently transform categorical variables into continuous numeric values. And third, in addition to the patient and embryo characteristics, we investigate the effect of individual physicians as a human factor on the pregnancy outcome. Finally, we apply Bayesian Networks to model the embryo growth process with the objective of blastocyst score prediction. We propose a novel approach to adjust the frequency estimates for parameter learning in conditional probability tables. The results of the experiments show that (i) the standard machine learning algorithms enable acceptable prediction of

implantation and blastocyst score and ii) the prediction performance can be improved by using the proposed techniques in this study. From the clinical perspective, our results have practical implications in reducing multiple pregnancies, preventing waste of embryos and cancelation of transfers.

ÖZET

TÜP BEBEK TEDAVİ SÜRECİNDE YAPAY ÖĞRENME YÖNTEMLERİ: İMPLANTASYON VE BLASTOSİST GELİŞİMİNİN KESTİRİMCİ MODELLENMESİ

Bu tezde tüp bebek tedavisinde klinik başarı oranlarının arttırılması için karar verme problemleri yapay öğrenme bakış açısı ile ele alınmıtır. İlk olarak, embriyo bazlı implantasyon tahmini için sınıflandırma tekniklerinin kapsamlı ve karşılaştırmalı bir analizi sunulmuştur. Aynı zamanda, özniteliklerin belirleyici etkileri değerlendirilmiş ve gereksiz değişkenler elenerek en iyi kestirim performansı oluşturan ideal öznitelik alt kümesi belirlenmiştir. Literatürde yer alan az sayıdaki ilgili çalışmada ifade edilenlerin aksine, başlangıç deneyleri sınıflandırıcı yöntemlerin tüp bebek tedavisinde potansiyel karar destek araçları olabileceğini göstermektedir. Çalışmanın devamında, metodolojik iyileştirmeler ya da verikümesinin bilgi içeriğinin genişletilmesi ile tahmin performansının arttırılması üzerinde yoğunlaşılmıştır. İlk olarak, dengesiz sınıf dağılımı problemi ele alınmış ve karar eşik değerinin optimize edilmesi ile öğrenme kümesinin tekrar örneklenmesi benzer sonuçlar oluşturmuştur. İkinci olarak, kategorik özniteliklerin sürekli sayısal değerlere dönüştürülmesi için frekans tabanlı bir kodlama yöntemi önerilmiştir. Üçüncü olarak, hasta ve embriyo özelliklerine ek olarak, doktorların deneyimlerinin tedavi sonucuna olan etkisi incelenmiştir. Son olarak, blastosist skoru tahmini için Bayes Ağlar yöntemi kullanılarak embriyo gelişim süreci modellenmiştir. Koşullu olasılık tablolarındaki parametrelerin daha iyi öğrenilebilmesi için yeni bir yöntem önerilmiştir. Deneylerde i)standard yapay öğrenme yöntemlerinin implantasyon ve blastosist skoru tahmininde kabul edilebilir başarı oranı elde ettiği ve ii)bu çalışmada önerilen yöntemler kullanılarak tahmin performansının arttırılabileceği görülmektedir. Bulgular klinik açıdan çoğul gebeliklerin azaltılmas, embriyo kayıplarının azaltılması

ve transfer iptallerinin engellenmesini sağlayacaktır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial neural networks
AUC	Area under the receiver operating characteristics curve
CBR	Case based reasoning
CPT	Conditional probability table
DET	Double embryo transfer
DFE	Discriminative frequency estimate
DT	Decision trees
eSET	Elective single embryo transfer
FE	Frequency estimate
FPR	False positive rate
FSH	Follicular stimulating hormone
HMM	Hidden markov models
ICSI	Intra-cytoplasmic sperm injection
IVF	In-vitro fertilization
kNN	k-Nearest neighbor
MLP	Multi-layer perceptron
OPU	Oocyte pick up
PCA	Principal component analysis
PGD	Pre-implantation genetic diagnosis
PGS	Pre-implantation genetic screening
RBF	Radial basis function
ROC	Receiver operating characteristics
SET	Single embryo transfer
SQL	Structured query language
SVM	Support vector machines
TAN	Tree augmented network
TET	Triple embryo transfer
TPR	True positive rate

1. INTRODUCTION

Infertility is defined as couple's biological inability to get pregnant after at least 12 months of regular, well-timed sexual intercourse without any birth control. It is reported that almost 10% of couples cannot have baby spontaneously. There may exist many medical disorders underlying infertility. Once the infertility factor of a couple is determined, an appropriate assisted reproduction treatment is used in order to conceive a successful pregnancy.

In-vitro fertilization (IVF) has been the most common infertility treatment method since 1978 [2]. IVF is a process which female germ cells (oocytes) are inseminated by the sperm under laboratory conditions. Development of fertilized oocytes (embryos) are observed for 2-6 days in laboratory and selected embryo(s) are transferred to the woman's womb at cleavage stage (day 2-3) or at blastocyst stage (day 5-6). Multiple embryo transfers increase pregnancy probability but also increase possible complications of multiple pregnancies for both mother and babies. Many researchers have been seeking various solutions to confidently implement single embryo transfers.

Approximately 40% of IVF treatment cycles result in successful pregnancies. The factors affecting the success of individual treatment cycles are mostly related to patient's response level and embryo viability. On the other hand, the overall success rates of IVF clinics depend on the medical equipment technology, treatment methods and personal experiences of clinicians and embryologists.

The complex structure of IVF process can be modeled using machine learning methods providing automated decision support to clinicians when necessary. On the contrary to the emergence and importance of decision support systems in IVF process, the related literature is limited. Artificial Neural Networks (ANN), Case-Based Reasoning System (CBR), Decision Trees (DT), Naive Bayes classifiers and logistic regression models are used as prediction methods in IVF treatment. However, the presented results are far from to be used in clinical practice as discussed in the pub-

lished studies. Furthermore, these studies represent a wide variety in dataset features, dataset size, outcome measures and performance criteria.

From the clinical perspective, the research on prediction systems in IVF treatment mostly focus on cycle based models handling the embryos as a cohort and predicting the outcome of the cycle as positive or negative. On the other hand, the most critical decisions such as transferring, further culturing and freezing are given for each individual embryo separately and this requires embryo based predictions.

Consequently, constructing reliable and practical embryo based prediction models in IVF treatment is still an open question as an interdisciplinary research interest. Non-automated analysis of various patient and embryo related parameters is difficult for clinicians in IVF domain. A computer assisted decision support system can automatically analyze large IVF databases, determine the relationships between predictor variables and outcome; and provide future predictions. Such a system would speed up the decision process, provide cost-efficiency preventing the waste of embryos and possibly improve the success of the treatment.

The most challenging problems of machine learning studies in medical domain are related to data retrieval. Unfortunately, there are no public IVF datasets to be used in machine learning experiments. When constructing a new dataset from an existing database, lack of necessary predictor variables in the database, missing or incorrect data records and security and privacy issues complicate the initial step of the research. As the next step, one needs to select the most appropriate machine learning methods for pre-processing of the data, classification and post-processing of the results. This step requires a good understanding of the underlying characteristics of the IVF domain as well as comprehension application of machine learning algorithms.

1.1. Research Overview

During this thesis period, we mainly concentrated on predictive modeling of embryo implantation and blastocyst development. After analyzing the limitations of ex-

isting machine learning applications in IVF process, we performed experiments to build up novel decision support systems as a benchmark study aiming to pave the way for further studies.

Our first research interest, predicting embryo implantation outcome consists of the following subtasks: construction of the dataset, application of state of the art classifiers comparatively, handling the constraints of the standard methods in order to improve the prediction performance and investigating the effect of the physicians experience as a human factor in success of IVF treatment.

- *Dataset construction and classifier selection:* The initial step of this thesis is collection of data and use of well known classification techniques for implantation prediction problem. To the best of our knowledge, this is the first embryo based dataset in IVF domain including both embryo morphological observations and patient and cycle characteristics. Each row feature vector in the dataset represents an individual embryo and the class label 1 and -1 indicates implantation and no-implantation, respectively. We have used the most popular representatives of different classifier categories because comparative analysis of diverse classifiers enables determination of the best fitting models in application domain. We have used ROC analysis for comparison and evaluation of classification performance. We have performed feature selection and feature extraction in order to reduce the computational cost in the rest of the experiments.
- *Handling the imbalanced class distribution:* The dataset represents an imbalanced distribution of class samples (11% positive class and 89% negative class). Sampling methods such as over-sampling and under-sampling have been proposed to balance the number of instances in the classes. On the other hand, we show that optimizing the threshold of classification produce results similar to re-sampling methods [3]. In addition, analysis of under-sampling experiments led to define sufficient size of embryo samples for implantation prediction that would reduce the effort spent for data collection in IVF domain.
- *Transformation of categorical variables in mixed IVF dataset:* The dataset we analyzed includes both categorical (infertility factor, treatment protocol etc.) and

continuous (e.g. age, hormone levels etc.) feature values. Transformation of categorical variables into numeric attributes is an important pre-processing stage for distance based algorithms such as Support Vector Machines (SVM), k-Nearest Neighbor (kNN) etc. affecting the performance of the classification. We have proposed a frequency based encoding technique for transformation of categorical variables [4]. Experimental results revealed that, the proposed technique significantly improved the performance of IVF implantation prediction compared to common binary encoding and expert judgement based transformation methods .

- *Analysis of effect of physician factor:* We have analyzed the effect of the experience level of individual physicians in success of embryo transfer in terms of pregnancy rate. We concluded that patient and embryo characteristics have greater impact in pregnancy rates. When these characteristics are compromised the level of physician experience may be a more determining factor [5].

The second research objective focuses on modeling embryo growth process from the oocyte collection (day 0) to blastocyst stage (day 5-6). After the analysis of different candidate models, we decided that Bayesian Networks best fitted the process characteristics. We have performed experiments for the two main subtasks:

- *Bayesian Networks for predicting blastocysts score:* Extended culture until the blastocyst stage (day 5/6) enables self-selection of the most developmentally competent embryos in IVF process since all the embryos cannot reach this stage. Delaying the transfer increase the implantation probability but also increase the risk of transfer cancelation if no blastocysts develops by day 5. We have used Bayesian Networks for predicting the blastocyst score in an embryo based dataset aiming to minimize waste of embryos due to developmental failure during extended culture [6].
- *Handling the problem of insufficient frequency estimates:* Learning Bayesian Networks from data requires network structure learning and parameter learning. In this part of the research, we have concentrated on parameter learning because of the dataset characteristics. Parameter learning in Bayesian networks is often based on Frequency Estimates which determines the conditional probabilities by

computing the frequencies of instances from the data. The main drawback of frequency estimate method in IVF dataset is related to the distribution of the training instances over data points where specific data points are represented infrequently. We propose a weighted nearest neighbor based approach to handle the problem of insufficient frequency estimates.

This research is mainly concentrated on predictive modeling of IVF treatment procedure as a novel application domain in machine learning community. The proposed modifications to standard machine learning algorithms produce enhanced prediction performance in IVF domain as well as presenting potential of generalization to other real world applications. We have used public datasets to validate our results obtained by the proposed method for adjusting the conditional probabilities in Bayesian Networks.

Since we obtained the dataset from a still growing and evolving database of Bahceci IVF Center, the dataset features and the number of instances changed during the experiments.

1.2. IVF Treatment Procedure and Success Criteria

The complete IVF procedure consists of controlling the follicular stimulation by external administration of hormones, aspirating oocytes from woman's ovaries (a.k.a. oocyte pick up - OPU), inseminating the oocytes with sperm cells in vitro, culture of embryos in the laboratory for 2-6 days and transferring the selected embryo(s) into the womb. This procedure is called an IVF cycle.

After 1992 the IVF process is combined with intra-cytoplasmic sperm injection (ICSI) method [7] which is the direct injection of a single sperm cell into the cytoplasm of the oocyte. ICSI has been an effective treatment for male infertility problems such as sperm defects or very few numbers of and/or immotile (motionless) sperms. Nowadays in most clinics ICSI has been a routine technique in IVF process.

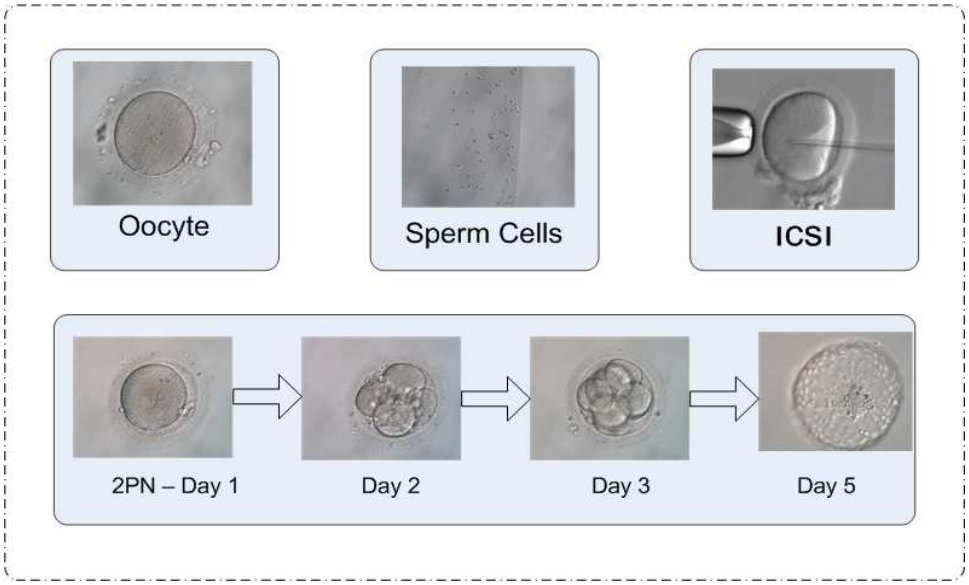


Figure 1.1. ICSI insemination and embryo growth day by day

After the oocytes have been inseminated with sperm cells, the embryos need to be continuously observed by embryologists. Gametes and embryos are kept in specially designed equipment which mimic the body conditions called incubator whenever they are exposed to laboratory conditions - outside the human body. The ICSI time is assumed as origin and embryo morphological parameters are manually recorded within certain intervals [8]. Figure 1.1 represents images to give emphasize on ICSI and embryo morphology.

The next stage is embryo transfer which has been performed by a gynecologist. In fresh IVF cycles, majority of embryo transfers are performed in day 2 or day 3 post ICSI. Selected embryo(s) have been transferred to the woman's uterus in IVF laboratories. The conventional and most common way of selecting high quality embryos is to inspect their morphologies. These morphological observations are evaluated by embryologists according to some pre-defined embryo scoring schemes [9].

In each IVF cycle, it is possible to obtain more than sufficient number of high quality germ cells or fertilized oocytes. In such cases patients are informed and offered to freeze high quality oocytes, sperm cells and embryos considering potential lack of patients' germ cells in the future. Freezing germ cells and embryos is a rapidly

developing technology allowing couples to have babies in advanced ages.

The main criteria for measuring IVF success rates are implantation rate and pregnancy rate.

Implantation Rate: Implantation is attachment of the embryo to the inner wall of the womb. A positive implantation is defined as visualization of the pregnancy sacs under ultrasound after 12 weeks of the embryo transfer. Implantation rate is an embryo based success measure:

$$\text{Implantation Rate} = \frac{\text{number of embryos implanted}}{\text{number of embryos transferred}} \tag{1.1}$$

Pregnancy Rate: Pregnancy is a positive implantation outcome regardless of number of embryos implanted. Pregnancy rate is a cycle-based success measure and defined as:

$$\text{Pregnancy Rate} = \frac{\text{number of positive outcomes}}{\text{number of transfer cycles}} \tag{1.2}$$

Multiple pregnancy and live birth rates are also important success criteria in IVF treatment. The major objective of the researches in IVF is to increase the implantation, pregnancy and live birth rates while reducing the numbr of multiple pregnancies. Increasing success rates depend on progress in treatment methods, medical equipment technology and critical decisions during the treatment. The critical decisions and the necessity of automated decision support in IVF process is discussed in the next section.

1.3. Decision Support in IVF

At each stage of IVF treatment patients need to know something about the outcome and physicians need to make decisions based on past experience or future predictions. Requirement for outcome prediction arises immediately at the beginning of the treatment. Prior to treatment, what patients want to know is "What is my chance of getting pregnant". This is one of the most difficult outcome prediction problems in IVF, since only age of patients, infertility factor and number of previous IVF attempts (if exists) is known at the beginning.

If a couple decides to start the treatment, helping physicians to select the most appropriate treatment method, i.e. type, duration and doses of stimulating medicine, may improve the success of outcome. While the treatment progresses, the number of prognostic variables increase leading more accurate outcome prediction but complicating the data analysis and inference, as well. After the oocyte collection, the most critical decisions are related to transfer, freezing and extended culture of embryos in the laboratory. Clinicians have to decide how many embryos, which ones and when to transfer, which ones and when to freeze and which ones to further culture.

A schematic representation of the standard IVF laboratory process is given in Fig. 1.2 [1]. This map demonstrates IVF treatment as a complex and costly process. Clinicians need to make critical decisions under uncertainty conditions. These critical decisions can be summarized as follows:

- Selection of the most appropriate treatment protocol for each individual patient.
- Decision of day of embryo transfer.
- Selection of embryos with highest implantation potential.
- Decision of number of embryos to be transferred.
- Identification of patients suitable for single embryo transfer.
- Decision of extended culture of embryos until the blastocyst stage.
- Determination of which embryos and when to freeze.

Decision support for treatment selection is a gynecological issue and out of scope of this study since we focus on the embryology side of IVF procedure. Decisions 2-5 are all related to implantation prediction of embryos. Figure 1.a points out this challenging decision process after day 2 and day 3 assessment of embryos. Figure 1.b represents the proposed machine learning based embryo implantation prediction system considering the afore mentioned quality factors and requirements of the embryo selection problem. Considering the decisions 6 and 7, we have analyzed the potential of predicting blastocyst score as the second research objective of this thesis.

The first step in a machine learning application is determination of the input predictor variables that affect the output. Medical literature represents various studies investigating the statistical relation between prognostic factors and IVF outcome. Physicians make the decisions considering these relationships and their prior experience. Increasing number of prognostic input variables complicates the data analysis process prior to critical decisions. Machine learning methods can be used as advanced prediction systems by analyzing large amount of data and providing reliable future predictions. However, there are very few studies handling the predictive models as a multidisciplinary research interest involving IVF and machine learning.

1.4. Literature Review

The relationships between embryo and patient characteristics and IVF treatment success rate have been investigated over the years and still attracting academicians as an emerging research field. Existing studies heavily focus on statistical relationships between clinical variables and pregnancy outcome [10, 11]. These studies provide valuable information for improving pregnancy rate. However, because of the difficulty faced in manual observation of multiple variables and examination of nonlinear correlations between features, IVF process requires more advanced data analysis and prediction models. On the contrary to the emergence and importance of intelligent decision support systems in IVF process, the related literature is limited.

Recent studies present applications of machine learning methods in IVF process

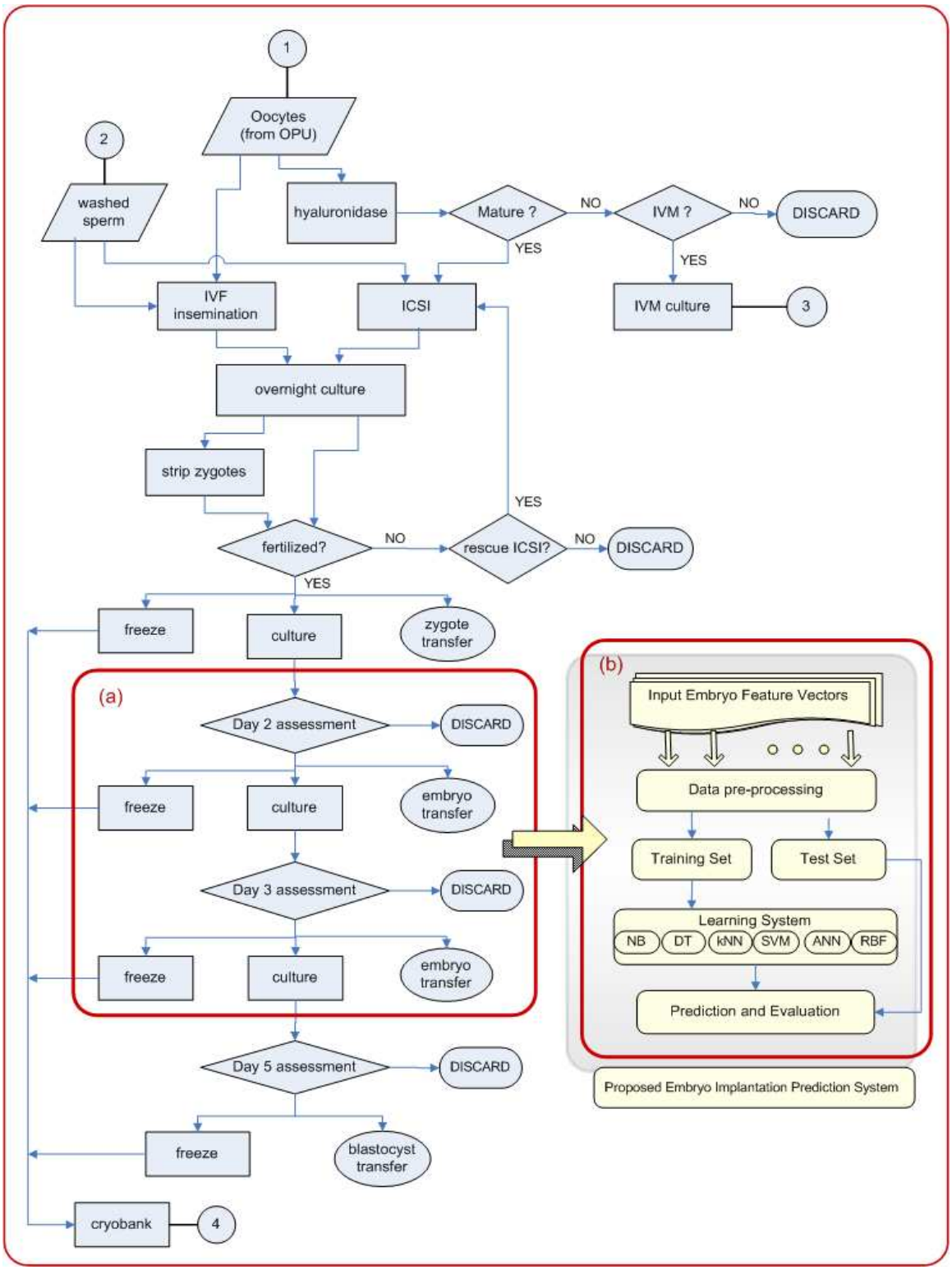


Figure 1.2. A map of the standard IVF laboratory process (from [1])

for different objectives. The majority of studies deal with prediction of implantation or pregnancy outcome which can be summarized in two groups. The first group of studies consider features related to characteristics of patients and embryo batches such as age, hormone levels, number of eggs, fertilization rate, number of embryos transferred etc. [12–15]. These studies presented contradictory results about the sufficiency of prediction performances.

Initially, Kaufmann et. al. constructed ANNs in cycle based prediction of pregnancy outcome using the variables of age, number of eggs recovered, number of embryos transferred and whether there was embryo freezing [12]. They have achieved a prediction accuracy of 59% and concluded that *"...the input information was not sufficient to characterize the outcome"*.

Trimarchi et. al. utilized C5.0 DT algorithm to retrospectively investigate the predictive power of the 100 parameters related to patient demographics, stimulation regime, response properties, oocyte and embryo characteristics [13]. They have collected records for each IVF cycle and predicted pregnancy outcome with 75% accuracy.

Jurisica et. al. propose a CBR system for two distinct purposes, first suggesting possible modifications to an IVF treatment plan in order to improve overall success rates, and second predicting pregnancy outcome [14]. The accuracy of outcome prediction is reported as 71% with input patient and cycle characteristics.

The most recent study on pregnancy outcome prediction presented that IVF cycle outcomes were predicted at 70% by four embryo cohort-specific variables which are total number of embryos, number of 8-cell embryos, percentage of cleavage arrest in the cohort and day 3 follicle stimulating hormone (FSH) level [15]. They were concluded that these cohort specific parameters were remarkably more informative than any measures of individual, transferred embryos.

The second group of studies predicting IVF outcome mostly concentrate on characteristics of individual embryos such as cell number, embryo grade etc. However, both

studies mentioned below consider cycles in which three embryos transferred, and either average the feature values over three embryos or include the features of all the three embryos in the same record.

Saith et. al. applied C4.5 class probability tree model in order to express relationships as simple rules of features characterizing as "take home baby" and "no take home baby" classes of embryo batches [16]. Fifty-three embryo, oocyte and follicular features were averaged over the three embryos in the batch and the relationship between features and outcome of transfer was analyzed. Only four of the 53 features (embryo grade, cell number, follicle size and follicular fluid volume) were identified as predictive. They have obtained 74% test accuracy and concluded that the results were satisfactory.

Morales et. al. propose a Bayesian classification system for embryo selection and reported an accuracy of 71% [17]. They consider transfer of embryo batches including 3 embryos and predict implantation outcome of the batch rather than individual embryos. Characteristics of the three embryos are included sequentially in the same data feature vector. In this case it is not possible to know which embryo of the batch is implanted and such an ambiguity challenges the reliability of the embryo selection mechanism.

Since there are no public IVF databases, all of the studies mentioned above perform experiments on different proprietary datasets. A direct comparison of reported results is not possible due to the varieties of outcome measure, data features, dataset sizes, methodologies and performance criteria. On the other hand, we conjecture that cycle level IVF outcome prediction does not exceed an upper limit of accuracy. Also, triple embryo batch based prediction mechanisms can not be applied to embryo selection problem because of the ambiguity of individual embryo implantation outcomes.

1.5. Thesis Outline

This dissertation is organized as follows:

Chapter 1 is the introductory part presenting research overview, explanation of the entire IVF process together with the requirements for automated decision support from clinical perspective and a literature review on machine learning applications in IVF domain.

We present the problem statement and relevant research questions in Chapter 2. We propose solutions for each research question in Chapter 3.

Chapter 4 presents the brief definitions of the machine learning algorithms as the methodology of our study. The experiments and results are given in Chapter 5.

Finally in Chapter 6, we provide an overall conclusion and discussion of the future research directions.

2. PROBLEM STATEMENT: BACKGROUND AND RESEARCH QUESTIONS

In this chapter we discuss our research questions with related problem statement and background. We mainly state five research questions with additional considerations and sub-questions.

2.1. Multiple Pregnancies and eSET

At each cycle of IVF treatment it is possible to obtain many embryos, but generally at most 3 highest quality embryos are transferred to the woman's uterus. Multiple embryo transfers increase pregnancy probability but also increase possible complications of multiple pregnancies [18–21].

Elective single embryo transfer (eSET) has been favored as a solution to IVF multiple pregnancy problem. However, applicability of eSET is limited due to the challenging tradeoff between increasing implantation rate and reducing multiple pregnancy rate. Reported success rates provide reasonable explanation for this tradeoff.

2.1.1. IVF Success Rates

National IVF success rates are reported by Society for Assisted Reproductive Technologies (SART) in USA [22] and by Human Fertility and Embryology Authority (HFEA) in UK [23] annually. Unfortunately, such a report is not provided in Turkey either by Ministry of Health or by any other organizations.

IVF success rates are evaluated for different age groups since age of the woman is an important factor on the outcome where increasing age reduces the potential of positive outcomes. Multiple embryo transfers are allowed in USA while the number of transfer embryos are restricted in UK and throughout the Europe. Reported success rates for the year 2008 in USA and in UK are summarized in Table 2.1. The presented

Table 2.1. Reported national success rates in USA and UK

	Age of woman					
	<35		35-37		38-40	
Success Criterion	USA	UK	USA	UK	USA	UK
Live births per cycle	41.3	33.3	31.1	27.3	22.2	19.4
Single live births	64.8	71.0	69.9	77.5	74.8	82.6
Live births with twins	33.3	28.7	281.	22.1	23.5	17.1
Live births with triplets	1.9	0.4	2.9	0.4	1.7	0.3
Pregnancy rate	47.6	-	38.0	-	30.3	-
Implantation rate	34.1	-	24.8	-	16.7	-
eSET rate	5.2	-	3.2	-	1.0	-
Avg. number of embryos transferred	2.2	-	2.4	-	2.7	-

success rates show the effect of number of embryos transferred in IVF outcome. Live births per cycle are higher in USA as well as multiple birth rates due to multiple embryo transfers.

In Turkey multiple embryo transfers has been a common procedure in IVF treatment until March 2010. Transfer of maximum of three embryos were allowed but in some conditions (more than two failed cycles, advanced maternal age (>38 years), PGD cases) four to five embryos could be transferred. Consequently, in a study investigating the global variations in the uptake of single embryo transfers between 2003-2005, the percentage of single embryo transfers in Turkey is reported as <10 which is one of the lowest rates over 31 countries [24].

2.1.2. IVF Legislation in Turkey

The previous legislation in Turkey provided higher pregnancy rates but higher multiple pregnancy rates as well. Ministry of Health published a new legislation in March 2010 [25]. The new regulations limit the number of transfer embryos to one in the first two cycles in women <35 years of age. A maximum two embryos is allowed to transfer in the third and subsequent cycles of women <35 years old and in all cycles

of women at 35 or older. These regulation is expected to reduce the complications of multiple pregnancies.

We conduct this research in collaboration with Bahceci IVF Center which is the largest IVF clinic in Turkey with the highest success rates. However, the number of single embryo transfers is still too low and hence improving the reliability of the eSET is crucial in Turkey and especially in Bahceci clinic.

2.1.3. Limitations of eSET

Despite the legislations, single embryo transfer is still accepted as a risk because of various domain related technical or economical reasons. Recently, van Peperstraten et. al. published a survey on perceived barriers to eSET among IVF professionals [26]. They have reported that, 47% of the IVF professionals refuse use of eSET associated with uncertainty about eSET technique or lack of prognostic factors and models to determine eSET candidates. Consequently, lack of a reliable eSET criteria is shown to be an important factor preventing clinical applicability of eSET.

Within the limits of available legislations, the decision for number of embryos to be transferred and selection of embryos depends on immediate analysis of available clinical records. This critical decision in IVF practice is usually based on a combination of clinical patient and embryo characteristics and embryologist's knowledge and experience. With increasing experience, such decisions become almost intuitive for most clinicians. However, the reliability of this intuitive approach is controversial because of the uncertainty in decision making process. Therefore, IVF experts need automated decision support tools in making the right decision on the number of embryos to transfer.

2.1.4. Implantation Prediction

In order to overcome the limitations of eSET, clinicians need reliable eSET criteria depending on two main issues: selection of the most viable embryos and identification of

patients suitable for eSET. These two issues should be processed as a single problem: predicting implantation outcome of individual embryos depending on both embryo morphological observations and patient and cycle characteristics.

2.2. Characteristics of IVF Data

We need to analyze the predictor factors characterizing the outcome of IVF treatment in order to provide a reliable prediction model.

2.2.1. Predictive Factors

Embryo morphological observations and patient related data have been widely investigated as predictor factors characterizing the IVF outcome as discussed in Section 1.4. The studies reporting lower prediction performance either question the sufficiency of information content of their datasets [12] or point out investigation of new predictor features as future work [17] since improving the information content of datasets provides better recognition performance in machine learning applications.

2.2.2. Imbalanced Class Distribution

As shown in Table 2.1, the implantation rate is in the range of [16.7, 34.1] for different age groups. This can be interpreted as the proportion of negative implantation outcomes dominate positive ones in IVF treatment. Therefore, any embryo based dataset would represent an imbalanced distribution of positive and negative class samples. In such cases, the problem of imbalance should be handled to overcome possible bias towards the majority class in the learning and prediction tasks.

2.2.3. Mixed Data Type

The prognostic factors in IVF procedure include both continuous (e.g. age) and categorical variables (e.g. infertility factor) [27]. Transformation of categorical variables into numeric values or discretization of continuous variables is crucial for the

specific classification algorithms. Defining the most proper method for transformation produce better prediction results. The mixed data type characteristics of the dataset has been another important challenge in our research.

2.2.4. On the Effect of Pre- or Post-Processing

Each real world application of standard machine learning algorithms require careful pre-processing of input data, necessary modifications to learning algorithms and post-processing of the results if necessary.

2.3. From Day 3 To Day 5 In The Laboratory: Blastocyst Stage Transfer

Embryo transfers can be performed at cleavage stage (Day 2-3) or at blastocyst stage (Day 5/6) after ICSI. We considered cleavage stage embryo transfers in the problem of implantation prediction. In the rest of the research, we focus on predicting the reproductive potentials of blastocyst stage embryos.

Extended culture until the blastocyst stage enables self-selection of the most developmentally competent embryos in IVF process since all the embryos cannot reach this stage. Delaying the transfer increase the implantation probability but also increase the risk of transfer cancelation if no blastocyst develops by Day 5. Consequently, prediction of blastocyst development is an important research question in IVF domain.

2.3.1. Prediction of Blastocyst Score

Transfer of blastocyst stage embryos at day 5 is thought to result in embryos with high implantation potential increasing implantation and pregnancy rates in IVF treatment. When equal number of embryos are transferred, it is suggested that the probability of live birth is significantly higher after blastocyst-stage embryo transfer at Day 5 as compared to cleavage-stage embryo transfer at Day 2 or Day 3 [28]. It is also recommended that in patients with a top-scoring blastocyst, transfer of a single blastocyst should be considered [29] preventing possible complications of multiple

pregnancies.

However, extended culture of IVF embryos may result in transfer cancelation if no blastocysts develops. Considering further culture of embryos until day 5 with the expectation of good quality blastocyst development, there is a tradeoff between the higher probability of implantation success and the risk of transfer cancelation. If one can predict whether blastocysts will develop or not, the risk of transfer cancelation can be minimized.

Recently, a cycle based model has been applied to predict blastocyst transfer cancelation [30]. In a cohort of at least 5 good quality embryos, the authors propose a model to predict if any blastocyst will develop or not. This model is useful in the sense of preventing transfer cancelation, but there are limitations related to requirements of the model since it can be applied to only specific cycles.

Considering the tradeoff between increasing pregnancy rate and possibility of transfer cancelation, clinicians need reliable models to predict blastocyst development for individual embryos. It is necessary to model the entire embryo growth process in order to determine relationships between daily morphological variations of embryos.

2.3.2. Embryo Growth Process

Figure 2.1 represents the developmental stages of IVF embryos day by day. The initial state is considered to be the ICSI insemination process. Fertilization check is performed at 16-18 hours after ICSI. Early cleavage morphology is observed at Day 1. Number of cells, nucleus characteristics, fragmentation rate, equality of blastomeres and appearance of cytoplasm is graded at Day 2 and Day 3. Finally, if the embryo is decided to be cultured until Day 5, the morphology of the blastocyst is evaluated using Gardner scoring system [31].

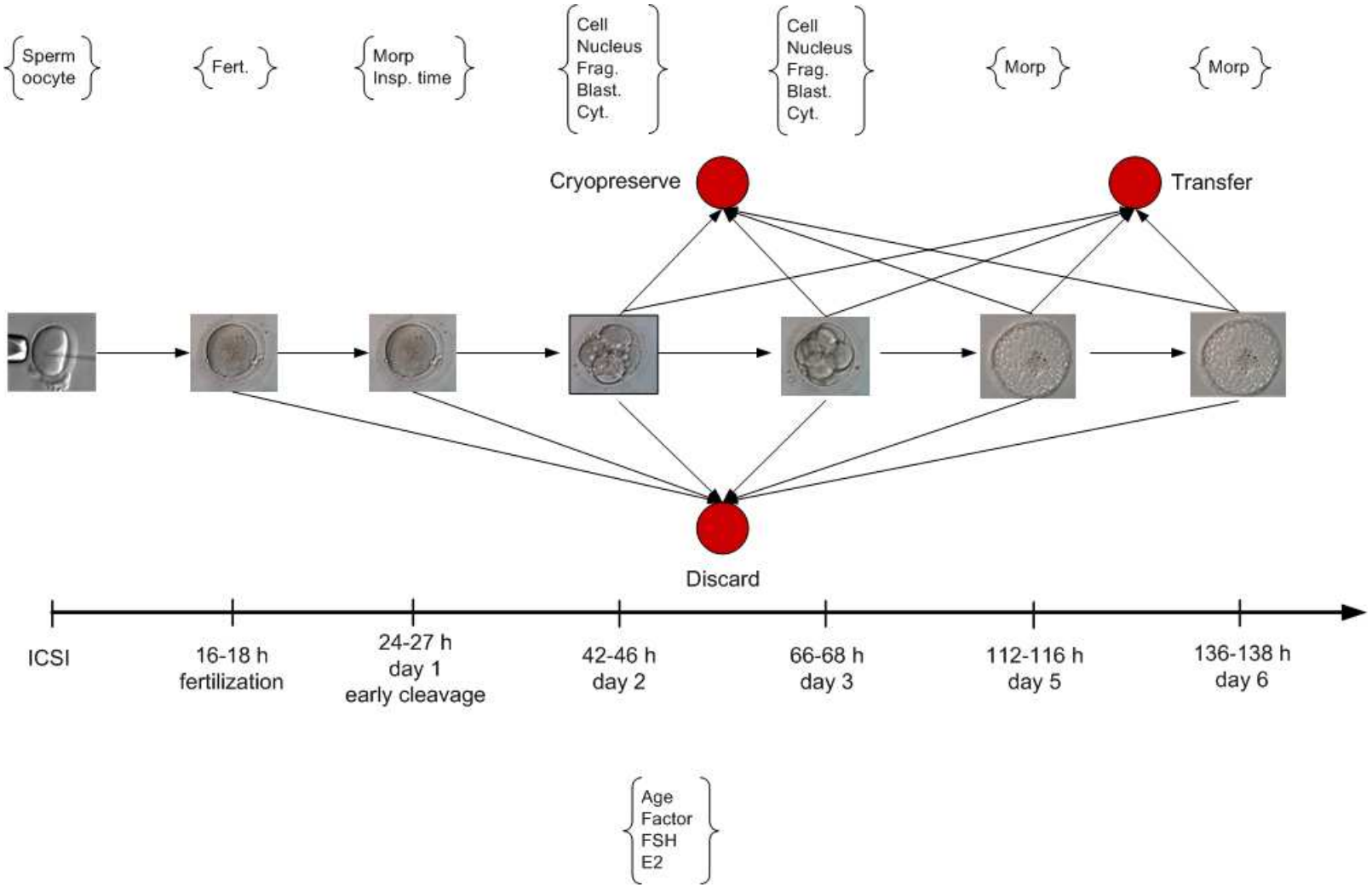


Figure 2.1. Embryo growth process with daily morphological observations and critical decisions

2.3.3. Blastocyst Morphology and Scoring System

Briefly, according to the Gardner's score blastocysts were graded based on the size:

- 1: early blastocyst, the blastocoel is less than half the volume of the embryo;
- 2: blastocyst, the blastocoel is greater than or equal to half of the volume of the embryo;
- 3: full blastocyst, the blastocoel completely fills the embryo;
- 4: expanded blastocyst, the blastocoel volume is larger than that of the early embryo and the zona is thinning;
- 5: hatching blastocyst, the trophectoderm has started to herniate through the zona; and
- 6: hatched blastocyst, the blastocyst has completely escaped from the zona.

For blastocysts graded as 3 to 6 (i.e., full blastocysts onward) the development of the inner cell mass (ICM) and trophectoderm can be assessed. The ICM grading is as follows:

- A: tightly packed, many cells;
- B: loosely grouped, several cells;
- C: very few cells.

The trophectoderm grading is as follows

- A: many cells forming a tightly knit epithelium;
- B: few cells;
- C: very few cells forming a loose epithelium.

A blastocyst with good morphology (usable for transfer) was defined as having a Gardners score $\geq$3AA. In order to predict the morphology of embryos at blastocyst stage, we need to specify the factors affecting the blastocyst development.

2.4. Research Questions

In this section we define our research questions based on the stated problems and relevant background presented in the previous section.

2.4.1. Research Question I: How can we construct an efficient embryo-based implantation prediction model?

Implantation prediction is a typical problem of decision making under uncertainty conditions because of the various factors affecting the outcome. We have stated the implantation prediction problem based on clinical requirements considering contributions and shortcomings of existing approaches. Rather than a comparison to previous studies, our objective is to build a novel applicable decision support system for all stages of IVF process by using advances in machine learning methods. Predicting implantation potential of individual embryos is the preliminary study of this research.

Any model on embryo selection is expected to provide a unique implantation outcome for each individual embryo. Because, the ideal case for eSET is transferring only one embryo with highest implantation potential and achieving positive pregnancy outcome. Considering the stated problem and relevant prior work in the literature, we performed experiments to construct an efficient embryo-based implantation prediction problem.

2.4.2. Research Question II: How can we enhance the methodologies to improve the prediction performance?

Our goal is to decide the best pre- and post-processing techniques to handle the imbalanced class distributions and mixed data type. We analyze the assumptions of the standard machine learning algorithms, compare the common pre- and post-processing techniques and propose modifications to improve the prediction performance in IVF domain.

2.4.3. Research Question III: How can we improve the information content of the IVF data?

In the first two research questions we deal with methodological convenience and algorithmic enhancements. Another aspect of our research is examining the predictor potentials of novel variables in IVF treatment.

In order to achieve improved information content, we analyze the association between available new features and IVF outcome. Specifically, we perform experiments to investigate the effect of human factor in success of IVF treatment.

2.4.4. Research Question IV: How can we model the embryo growth process?

In this part of our research, we investigate the time dependent development of IVF embryos and we apply Bayesian networks for predicting blastocyst score at Day 5 by modeling morphological evolution of IVF embryos.

2.4.5. Research Question V: How can we adjust the model parameters for prediction of blastocyst score when learning from data?

The IVF dataset that we used in experiments include thousands of embryo records and hence can be considered as a sufficiently large database. However, it is necessary to analyze if the observed frequencies are optimal.

3. PROPOSED SOLUTION

We outline our proposed solutions for each research question to be a base for the methodology and experiments.

3.1. Implantation Prediction as a Supervised Binary Classification Problem: Research Question I

A learning based predictor model that makes use of artificial intelligence (AI) notion can automatically analyze large medical databases to train predictor models and provide future implications. Specifically for the implantation prediction problem, these models can be used to predict the implantation outcome of embryos when relevant prognostic features are supplied as model inputs.

Quality of an intelligent learning based system depends on three main factors. First, construction of a comprehensive dataset that represents the underlying characteristics of the application domain enables accurately learning the relations between input and output. Second factor is, selection of best fitting model(s) for the specific domain together with unbiased training and testing strategies that avoid the sampling and learning bias. Third factor is careful application of the model specific pre-processing techniques and necessary algorithmic modifications to enhance the prediction performance.

Initially, we have constructed a dataset from an existing IVF database that forms a base for application of predictor models. We consider embryo-based prediction which is sufficient for reliable embryo selection. Each embryo is represented as a data feature vector including 18 clinical variables and a class label: +1 for implantation and *-1* for no-implantation respectively.

Dataset construction is followed by data pre-processing and classification stages. Data pre-processing includes eliminating samples with missing values and classifier

specific data normalization schemes.

3.2. Handling the Imbalance Problem and Mixed Data Type in IVF Dataset: Research Question II

The results of initial experiments on implantation prediction motivated us to improve the performance of classification. There are two ways to improve the performance of a classification task: to improve the algorithms to better fit the problem or to improve the information content of the data. Regarding our second research question, we performed experiments to improve the algorithms to better handle the imbalance and mixed data type problems.

Learning from imbalanced datasets has been an important research interest in the last decade [32, 33]. Various sampling strategies have been proposed to deal with the problem of imbalance [34–36]. On the other hand, recent studies show that adjusting the decision threshold of classifiers produce similar results with artificially changing the distribution of the instances in the training set [37, 38]. We apply under- and over-sampling strategies to re-balance the dataset and adjust the decision threshold to improve the classification results.

Analysis and pre-processing of mixed datasets including a combination of continuous and categorical variables is investigated widely [39–42]. In this research, we analyze the performance of implantation prediction on mixed IVF dataset using SVM method. We propose a frequency based encoding technique for transformation of categorical variables.

3.3. Investigating the Effect of Physician Factor in Success of Treatment: Research Question III

Our dataset includes only patient and embryo related variables. However, the manipulations associated with embryo transfer are also critical. The degree of difficulty during the transfer and the influence of the manipulating physician can also be

investigated as predictor factors affecting the outcome [43–47].

The former has been characterized by the type of catheter [45] as well as the presence of blood or mucus [46], and also the time spent for the procedure [47]. It has been shown that difficult transfers are associated with a reduced implantation rate [43, 44].

The latter, the physicians' impact on the success of treatment, is less certain as conflicting results have been reported [48–51]. One probability for such uncertainty may be due to the variation of the prior experience of physicians performing the transfer. Accordingly, the number of transfers varies to a great extent between the performers in the studies claiming difference among physicians [49, 50]. When the pregnancy rates have been shown as not dependent on individual performers [51], the physicians had been observed to have prior experience.

In the related experiments, we aim to analyze whether the prior experience levels of the physicians performing the embryo transfer influenced the outcome of the cycle. In order to analyze the differences between pregnancy rates of individual physicians, the statistical tests have been conducted to compare the pregnancy rate of each physician to the highest pregnancy rate.

3.4. Bayesian Networks for Modeling the Developmental Stages of IVF Embryos: Research Question IV

In this research, we apply Bayesian networks for modeling morphological evolution of IVF embryos and predicting blastocyst development. Bayesian network classifiers have been popular tools for medical decision support systems in the last decade [52, 53]. Specific applications include bypass surgery survival prediction [54], ovarian cancer diagnosis [55], diagnosis of female urinary incontinence [56], diagnosis and treatment of ventilator-associated pneumonia [57] etc.

The visualization of statistical cause-effect relationships in a network structure

makes the Bayesian networks easy to understand and apply in medical applications. The components of the Bayesian networks, i.e. nodes, arcs and conditional probabilities correspond to prognostic variables, dependencies and statistical inference, respectively. Such a model is useful especially when we need to know the underlying reason for the prediction outcome rather than a black-box model in which the explanation for the prediction is difficult to understand.

In our case, we have considered prediction of blastocyst score as a binary classification problem to discriminate blastocysts into two classes as potential high quality or low quality ones.

3.5. Clustering the Infrequent Combinations of Embryo Features in CPTs: Research Question V

In practice, generally the components of the Bayesian networks are unknown and must be inferred from the data. Learning a Bayesian network form data involves two subtasks, structure learning, which is necessary to identify the topology of the network, and parameter learning, that identifies the statistical parameters (conditional probability table (CPT)) for a given network topology.

Most studies concentrate on structure learning which is a complex procedure. Learning the parameters in conditional probability tables is recognized as a trivial task based on frequency counts of data points when the observed frequencies are optimal in a sufficiently large database [58].

The morphological embryo variables are categorical values including too many categories (and thus the conditional probability table is large) and there are few data samples to represent certain combinations of feature values. In such cases, the learning may be less than optimal, and it may be necessary to find another way of estimating the probability tables.

We consider the problem of limited number of samples to represent real condi-

tional probabilities as partially insufficient frequency estimates. We propose a weighted Nearest Neighbor approach to optimize the conditional probabilities to handle the insufficiency of parameter learning in Bayesian Networks.

4. METHODOLOGY

In this chapter, we provide the theoretical background of the statistical and machine learning techniques that we used in our experiments. We also discuss the relevance of the selected methods to our research questions regarding the characteristics of the IVF domain.

4.1. Experimental Design

The clinical studies can be categorized as retrospective and prospective according to data collection method and occurrence of events of interest. The definitions of the terms 'retrospective' and 'prospective' are given [59] as:

Retrospective: "All events of interest have already occurred and data are generated from historical records and from recall."

Prospective: "Data collection and the events of interest occur after individuals are enrolled (e.g. clinical trials and cohort studies)."

In our research, we mainly use retrospective data consisting of the completed cycles that we know the pregnancy outcome. Retrospective design provides cost and time efficiency because we use the available information. However, one major drawback of retrospective data collection is missing or erroneous variables in the database.

Prospective studies provide more robust, consistent and reliable results avoiding the potential biases in the historical data. Prospective validation of a prediction model in medical domain is necessary. However, considering the implantation prediction, yet it is not feasible for the clinicians to perform embryo transfers according to the decision of a machine learning system. Therefore, we used a semi-prospective approach [60] for validation of our retrospective results.

In the semi-prospective study we asked embryologists to predict the implantation outcome of embryos that were going to be transferred soon. We took the majority decision as the 'expert judgement'. We also simultaneously used our model to make future predictions. After 12 weeks of embryo transfer we compared expert judgement, and model predictions with the actual outcomes.

4.2. Data Collection

Because of social, ethical and financial reasons some legislative rules have been defined for assisted reproduction process in every country. Usually, the restrictions apply for donation, embryo manipulation, number of embryos to be transferred in each cycle etc.

Besides the legal procedures in countries, every IVF clinic apply different technologies and methodologies in practice even if they are in the same country. Because of this variety, each clinic has distinctive IVF databases. In this research, we will analyze the IVF procedure and related database of Bahceci Women Health Care and IVF Center at Istanbul.

Bahceci Women Health Care and IVF Center is the largest IVF center in Turkey with approximately 3000 patients' for each year. The overall patient-based pregnancy success rate is reported as more than 75%. Since some of the patients cannot be pregnant in first cycle, cycle-based pregnancy success varies between 45% and 55%.

Since 2004, patient, cycle and embryo related data is recorded and stored in Bahceci clinic. The resulting database presents an opportunity for machine learning studies. Before 2007, an excel based database existed with 9294 IVF cycles. This database includes both patient and embryo related data. Since beginning of 2007, an SQL-based relational database is used for recording relevant data of more than 3000 cycles. There are some differences between these two databases such as embryo grading strategy and IVF process time records.

4.2.1. Dataset

The dataset we used in this research has been constructed from a database which contained information on cycles that has been performed at the Bahceci Clinic in Istanbul from January 2007 through August 2008. Since the beginning of 2007, embryology laboratory of the hospital utilized a well-designed (Structured Query Language) SQL-based relational database for recording patient and embryo related data. The embryo based dataset used in the present study was obtained by performing SQL queries on this database.

The limited number of single embryo transfers performed in the Bahceci Clinic entail investigation of multiple embryo transfers for individual embryo implantation. In that case, it is necessary to consider arguments about the dependency of embryo implantations when transferring more than one embryo in an IVF cycle. In a recent study, Matorras et al. used a collaborative model for predicting the pregnancy rate and concluded that the implantation of one embryo is facilitated by the implantation of other embryo(s) [61]. In contrast, in another study no evidence was found for dependency of embryo implantations [62] and this fact was confirmed by [63]. Therefore, based on the assumption that embryos implant independent of each other, in addition to single embryo transfers, each embryo in the multiple embryo transfers was also represented as an individual record when the exact implantation outcomes were known.

Similar to existing studies in the literature [11, 64], implantation outcome of individual embryos was determined by assessing cycles with 100% implantation (i.e. number of sacs visualized was equal to the number of embryos transferred) and cycles with 0% implantation (i.e. negative implantation outcome). A total of 3898 embryo records were collected in this manner. Cycles with monozygotic twin pregnancies and samples with missing feature values were excluded from the experiments. The final dataset comprised 2453 embryos in which 273 embryos had proven positive implantation and 1870 embryos had proven negative implantation.

The overall implantation rate during the study period is reported as 25.8%

Bahceci Clinic. However, the implantation rate in the assessed dataset is 11% because of exclusion of pregnancies where some but not all the embryos implanted.

Dataset features and data types are given in Table 4.1. The features have been selected depending on experiences of senior embryologists in [47] and related studies in the literature [17]. The IVF dataset includes 2453 fresh, non-donor in-vitro human embryos transferred in day 2 or day 3 after Intra-Cytoplasmic Sperm Injection (ICSI) and each embryo data vector is represented by 12 feature values. There are two classes of embryos labeled as 1 and -1 indicating implantation and no-implantation, respectively.

4.3. Feature Selection

In dataset construction stage, we have included potential predictor features existing in the database. Each embryo was initially described with a vector of 18 feature values. However, all features may not be necessarily relevant to implantation outcome. In some cases, a reduced feature subset would better represent the information content of the underlying dataset and overcome "curse of dimensionality" in learning phase.

The aim of the feature selection is to find the k of the d dimensions [65]. Reducing the number of input features by eliminating the redundant variables is expected to:

- Improve the performance of prediction,
- Reduce the computational complexity of the learning algorithms,
- Prevent storage of unnecessary medical data,
- Provide better understanding of the underlying process, and
- Simplifies the utilization of the model in the clinical routine.

In our dataset, there are 2^{18} possible subsets of 18 input features. Testing all the subsets is not feasible computationally. In order to reduce the search space we need to apply some heuristics such as Information Gain feature weighting approach.

Table 4.1. Selected dataset features for each embryo feature vector

Dataset Features	**Data Type**
Patient Characteristics	
Woman age	Numerical
Primary or secondary infertility	Categorical
Clinical Diagnosis and Treatment Protocol	
Infertility factor	Categorical
Treatment protocol	Categorical
Duration of stimulation	Numerical
Follicular stimulating hormone dosage	Numerical
Peak Estradiol level	Numerical
Endometrium thickness	Numerical
Sperm quality	Categorical
Embryo Related Data	
Early cleavage morphology	Categorical
Early cleavage time	Numerical
Transfer day	Categorical
Number of cells	Numerical
Nucleus characteristics	Categorical
Fragmentation	Categorical
Blastomeres	Categorical
Cytoplasm	Categorical
Thickness zona pellucida	Categorical

4.3.1. Information Gain Feature Weighting

Information Gain represents the average amount of information about the class value C contained in the feature value F [66]. Information Gain is also known as mutual information between F and C.

$$InfoGain(F) = I(C, F) = H(C) - H(C|F) \tag{4.1}$$

where

$$H(C) = \Sigma_i P(C_i) log_2 P(C_i) \tag{4.2}$$

is the Shannon's entropy and

$$H(C|F) = -\Sigma_j P(F_j) \Sigma_i (C_i|F_j) log_2 P(C_i|F_j) \tag{4.3}$$

Higher Information Gain means higher predictor effect of the feature individually. The Information Gain values of features provide reasonable knowledge to reduce the search space for feature subset selection.

4.3.2. Forward Subset Selection

The predictive value of each input variable in classification has been examined by using the forward feature selection approach. In forward selection, classification starts with the single feature F_1 that has the highest rank. Then, other features are added one by one according to decreasing order of estimated ranks. The subset leading to best performance was selected and utilized in the rest of the experiments.

Let S be the selected subset of features and $E(S)$ be the error of classification using the features in S. For each feature F_i, $E(S \cup F_i)$ is calculated and F_i is added to the subset if it decrease the error.

add F_i to S if $E(S \cup F_i) < E(S)$

4.3.3. Filter Approach

The features are filtered according to the estimated Information Gain rankings. The features with an Information Gain value less than a pre-defined threshold are selected as the input parameters. For example, the threshold can be defined as the average of the Information Gain of all of the features, $\mu_{IG}(F)$. Then,

add F_i to S if $InfoGain(F_i) < \mu_{IG}(F)$

4.4. Feature Extraction

Feature extraction transforms the data in the D-dimensional space onto a space of d-dimensions, $(d \leq D)$, by linear or non-linear projection. The main linear approach for feature extraction is Principal Component Analysis (PCA) [67]. PCA performs a linear mapping of the original data to a new feature space where the derived features are uncorrelated.

In our case, the aim of feature extraction is to obtain uncorrelated features. Therefore, it is necessary to analyze the correlations of the input features prior to feature extraction. This will provide an understanding of the dependency structure of the dataset.

4.4.1. Dependency and Correlation Analysis for Features

The most common measure of the relationship between the random variables is the correlation coefficient which is derived from the covariance and variances of the random variables. Correlation analysis is useful for identifying the pairwise feature inter-correlations within a dataset.

The dependency structure of the data will be specified in terms of correlation ma-

trix which is a square symmetric matrix containing the correlation coefficients between each pair of input features.

4.4.2. Principal Component Analysis

PCA transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The principal components are linear combinations of original features. Each attribute is multiplied by a coefficient, where these coefficients correspond to the elements of the principal eigenvectors.

The mathematical technique used in PCA is called eigen analysis. A solution for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products is carried out. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component.

4.5. Classification

We have used Naive Bayes classifier (NB), k-Nearest Neighbor (kNN), Decision Tree (DT), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) as predictor models [65]. Comparative analysis of diverse classifiers enables determination of the best fitting models in IVF domain. We have chosen these classifiers because, we believe that the most popular representatives of diverse algorithms (statistical classifiers, decision tree approaches, neural networks, support vector machines and nearest neighbor methods) are included [68, 69]. We do not repeat the formulations of the selected classifiers here since they are well-known methods to machine learning community and we have performed only comparison experiments initially. On the other hand, Naive Bayes and SVM classifiers are further investigated, therefore a brief definition for these classifiers is presented.

Furthermore, we have exhaustively investigated and used Bayesian Networks for modeling embryo growth process. The formulation of Bayesian Networks is also summarized.

4.5.1. Naive Bayes

The aim of the classification in implantation prediction is to discriminate the embryo samples as 'implants' and 'not-implants'. Positive class (C_{+1}) and negative class (C_{-1}) denote implantation and no-implantation, respectively.

Nave Bayes classifier computes the class posterior probabilities, $P(C_i|x)$ of input embryo data (x) for both negative and positive implantation classes.

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)} \tag{4.4}$$

In case of binary classification, the default decision threshold was 0.5 and the embryo was decided to belong to the class with the highest posterior probability.

$$\text{choose} \begin{cases} C_{+1} & \text{if } P(C_{+1}|x) \geq 0.5; \\ C_{-1} & \text{otherwise.} \end{cases}$$

4.5.2. Support Vector Machines

Given a set of training data pairs (x_i, y_i), $y_i \in \{+1, -1\}$, the aim of the SVM classifier is to estimate a decision function by constructing the optimal separating hyperplane in the feature space [70]. The key idea of SVM is to map the original input space into a higher dimensional feature space using kernel functions. Final decision function is in the form:

$$f(x) = \left(\sum_i \alpha_i y_i K(x_i \cdot x) + b \right) \tag{4.5}$$

where $K(x_i \cdot x)$ is the Kernel transformation. The most popular kernel functions are:

Linear: $K(x_i \cdot x) = x_i^T x$

Polynomial of degree p: $K(x_i \cdot x) = (1 + x_i^T x)^p$

Radial Basis Function: $K(x_i \cdot x) = \left[-\frac{\|(x_i - x)\|^2}{\sigma^2}\right]$

The optimum Kernel function and related parameters should be selected in the training phase when using SVM classification.

A penalty term C is defined as an upperbound on the Lagrange multipliers α_i trading off the complexity of the algorithm and misclassification.

$$0 \leq \alpha_i \leq C, \forall i \tag{4.6}$$

A higher C minimize the misclassification but may also lead overfitting of the model. Therefore the value of C and needs to be tuned in the training phase in addition to Kernel parameters.

Finally, the class of an instance is decided according to the sign of the decision function:

$$\text{choose} \begin{cases} C_{+1} & \text{if } f(x) \geq 0; \\ C_{-1} & \text{otherwise.} \end{cases}$$

SVM computes the distances of instances to the separating hyperplane in the new input space. This computation is based on assumption of continuous numerical variables. However, the dataset may include categorical features as in our IVF dataset.

In that case it is crucial to transform the categorical variables to continuous numerical values. Transformation methods that we use in our experiments are summarized in Section 4.6 as a pre-processing step.

4.5.3. Model Selection for Embryo Growth Process

Initially we have investigated Hidden Markov Models (HMM) [71] as a candidate model for modeling blastocyst development. In a stochastic process with the Markov property, given the present state of the system, its future and past are independent. More precisely, the observation in any state depends only on previous state, not on any other past states (Equation 4.7).

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...] = P[q_t = S_j | q_{t-1} = S_i] \tag{4.7}$$

Assuming the time instants are the days associated with morphological states of embryos, we concluded that the characteristics of the embryo growth process does not meet the basic stochastic assumption of the Markov process. Because, in published studies it is reported that the blastocyst morphology at day 5 depends not only on day 3 but on day 1 and day 2 morphology [72].

Actually, the researchers are still investigating the statistical properties of embryo morphological evolution and dependencies between embryo development and patient characteristics. Literature presents conflicting results about predictor factors and their correlations. Therefore, as a starting point, we need to construct a model to analyze all available features and their statistical relations to blastocyst morphology.

We decided to investigate the potential of Bayesian networks in analyzing the statistical relationships between sequential observations of embryo morphology.

4.5.4. Bayesian Networks

A Bayesian network is a directed acyclic graphical model that encodes probabilistic relationships among variables of interest [73].

A brief definition of Bayesian networks and Bayesian network classifiers [74] is given below:

Bayesian networks allow efficient representation of the joint probability distribution over a set of random variables. The network structure is used to characterize a probability distribution for each node depending on its parents. And then, posterior probabilities are computed in the form of local conditional distributions.

A Bayesian network is represented by $B = \langle G, \Theta \rangle$, where G is a directed acyclic graph. The nodes of the graph correspond to the random variables $X_1, ...X_n$ which are the dataset features and edges represent direct dependencies between the associated variables. The graph G encodes the independence assumption where each variable X_i is independent of its nondescendants given its parents Π_{X_i} in G. The second component Θ represent the conditional probability distribution that quantifies the dependency between the nodes.

A Bayesian network defines a unique joint probability distribution over the set of random variables X_i in the network given by:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | \Pi_{X_i}) \tag{4.8}$$

where Π_{X_i} denotes the set of parents of X_i in the network.

In practice, generally the components of the Bayesian networks are unknown

and must be inferred from the data. Learning a Bayesian network from data involves two subtasks, structure learning, which is necessary to identify the topology of the network, and parameter learning, that identifies the statistical parameters (conditional probabilities) for a given network topology.

Most studies concentrate on structure learning which is a complex procedure when there are lots of inputs feature [54, 58, 75]. Learning the parameters in conditional probability tables is recognized as a trivial task based on frequency counts of data points when the observed frequencies are optimal in a sufficiently large database [58]. Here, we review the main approaches for construction of the network structure and estimation of parameters when learning Bayesian networks from data.

4.5.4.1. Structure Learning. Structure learning is a search for encoding appropriate dependencies between the features of a given a dataset. It has been argued that Bayesian network structure learners are computationally expensive requiring an exponential number of conditional independence tests [58]. There are two main approaches to learn the network structure from data efficiently reducing the search space: constraint based methods and methods that maximize a selected score.

Simple learning algorithm (SLA) [58] and three-phase dependency analysis (TPDA) [58] are examples of constraint based methods that make use of information theory concept in order to reduce the computational complexity of the structure learning procedure. Reiz and Csato also propose a mutual information based approach where direct causal relations encoded by the BN are interpreted as the maximum of conditional mutual information between nodes [54].

The algorithms based on a scoring function attempt to find a graph that maximizes the selected score, which evaluates how well a given network matches the data. Different learning algorithms can be obtained depending on the definitions of the scoring function and on the search procedure used. Meloni *et al.* propose a variation of standard search-and-score approach that computes a square matrix containing the mu-

tual information among all pairs of variables [75]. The matrix is binarized to find what relationships must be prevented. This approach prevents the inference of too many connections.

Furthermore, there are well-known simple Bayesian network classifiers with highly constrained dependency structures: Naive Bayesian network assuming mutual independence of the feature variables given the class variable and Tree Augmented Network (TAN) representing a tree-like dependency structure over the feature variables [76].

In our experiments, we construct a constraint based Naive Bayesian network structure using mutual information between nodes.

4.5.4.2. Parameter Learning. Parameter learning in Bayesian networks is often based on Frequency Estimates (FE) which determines the conditional probabilities by computing the frequencies of instances from the data. The FE method is efficient since it counts each data point in the training set only once. The parameters estimated using FE method maximize the likelihood of the model given the data and thus FE is known as a generative learning method [77].

The relative frequencies in the CPT are obtained as follows:

$$\hat{P}(X_i = x|\Pi_{X_i} = \vec{u}) = \frac{count(X_i = x, \Pi_{X_i} = \vec{u})}{count(\Pi_{X_i} = \vec{u})} \tag{4.9}$$

In our case, X_i denotes the class label as the child node that is the blastocyst score and $\vec{u}$ denotes a vector of parent nodes Π_{X_i} representing the predictor factors affecting the blastocyst score.

The classification capability of FE method is argued because of the generative property. Grainer and Zhou proposed a gradient descent based discriminative parameter learning method, ELR, that significantly outperforms FE method with a high

computational cost [78].

A Discriminative Frequency Estimate (DFE) is proposed to maximize the generalization accuracy of classification rather than likelihood [77]. The authors compared the DFE and FE methods based on Naive Bayesian network structure and showed that DFE significantly improve the performance of classification in terms of accuracy. However, it has been widely accepted that accuracy is not an appropriate performance measure especially for imbalanced datasets. On the other hand, the training time of DFE method is significantly higher than FE method. Consequently, an efficient and effective method for parameter learning in Bayesian networks is still an open question.

We propose a method for parameter learning from data taking advantage of efficient FE method and handling the insufficiencies in the data.

4.5.5. Parameter Optimization

The performance of classification is influenced by selection of the classifier specific model parameters. We have applied a grid search approach to find the optimum parameters for the classifiers. Basically, we tested possible combinations of parameters using cross validation in the training phase and decided the values with best performance to be the model parameters in the testing phase.

Concerning the classifiers we used in our experiments, the model parameters include: number of neighbors (k) in k-NN; cost and kernel parameters in SVM; standard deviation and number of clusters in RBF; number of hidden layers, number of hidden units in each hidden layer, learning rate, momentum and number of epochs in MLP classification.

4.6. Pre-processing

4.6.1. Sampling Imbalanced Data

A common approach to overcome the problem of imbalance is to re-balance the datasets artificially. Two main sampling strategies are over-sampling that replicates instances from the minority class [35] and under-sampling where some of the instances in the majority class is removed [34].

4.6.2. Transformation of Categorical Variables

Performance of distance based classifiers, such as SVM, depends on accurate transformation of categorical variables into numeric data. SVM requires each data sample to be represented as a feature vector of real numbers [79]. Therefore, categorical features should be converted into numeric values prior to classification. After transformation of categorical variables, the input data were normalized to 0 mean and standard deviation of 1.

The aim of data type transformation is to preserve the information content of the original dataset while adapting the input data to a particular analysis tool. We use binary encoding in the initial experiments of SVM classification and propose a frequency based encoding technique for better transformation.

4.6.2.1. Binary Encoding. Binary encoding maps categorical variables to higher dimensional features representing equal Euclidean distances between categories and has been applied as a common pre-processing stage for SVM applications [79, 80].

For a particular categorical variable including N categories, each category is represented by a sequence of N bits. The i^{th} bit corresponding to original category is set to 1 and the others are set to 0. For example, the treatment protocol feature in IVF dataset includes eight categories. When binary encoding is applied, the categories

1,2...8 correspond to 00000001, 00000010... 10000000 respectively. In this case the Euclidean distance between each category is equal, however, this may not be the actual case. Also, the input dimensionality is increased by adding dummy variables that may yield to "curse of dimensionality" in learning phase [65, 81].

4.6.2.2. Proposed Frequency Based Encoding Technique. The literature present variances of binary encoding, frequency based and expert judgement approaches for transformation of categorical variables. However, comparative analysis of these methods is limited and also, to the best of our knowledge, there is not a generalized frequency based encoding scheme.

Johannson, et al., deal with visualization of mixed datasets and propose interactive quantization of categorical variables that incorporates information about relationships among continuous variables as well as makes use of the domain knowledge of the data analyst [82]. A Simple Correspondence Analysis (SCA) has been applied based on the frequencies of categories in the dataset.

A frequency based encoding scheme has previously been proposed as a data transformation technique for car injury prediction [83]. In this research, we propose a new frequency based transformation method for continuous numerical representation of categorical variables in mixed IVF data. The new numerical values are derived from the relative frequencies of categorical codes among both positive and negative implantation classes as defined below:

$$x_{ik} = P(C_p)_{ik} - P(C_n)_{ik}$$

where,

- k_{ik} is the new numerical value of categorical code x_i originally in code k;
- $P(C_p)_{ik}$ is the frequency of categorical code k in positive implantation class C_p, and
- $P(C_n)_{ik}$ is the frequency of categorical code k in negative implantation class C_n.

The basic idea behind this transformation is to reflect the effect of categorical code on implantation outcome. The frequency of any categorical code in positive class is assumed to have positive effect while the occurrence in negative class is considered as negative effect. Hence, the new numerical value of a categorical code is defined as the difference between frequencies in positive and negative classes in the range of [-1,1].

4.7. Training and Testing Strategies

4.7.1. Two-third One-Third Split

Two-thirds of the dataset was randomly selected for establishing a predictor model and the remaining one-third was utilized for testing. This random splitting was performed considering stratification principle in order to ensure that the proportions of implanted and not-implanted embryos were the same in both training and test sets as in the original dataset. For each classifier, the model parameters were optimized on the 2/3 dataset using 10 fold cross validation strategy.

After selecting the best parameters, the trained model was assessed on the separate 1/3 dataset to predict the class labels of the previously unseen data samples. Finally, the predictions were compared to the actual implantation outcomes in order to evaluate the performance of the classification model. The random train set and test set partitioning was also repeated 10 times in order to avoid the sampling bias. The reported results were the mean values of these 10 experiments.

4.7.2. 10-fold Cross Validation

We have used 10-fold cross-validation for parameter optimization on the training set. In 10-fold cross-validation, dataset is divided into 10 equal subsets, 9 subsets were used for training and 1 subset was used for testing. This task has been repeated 10 folds to ensure that each data sample is used for training and testing.

4.8. Performance Evaluation

Reliable evaluation of prediction results is crucial for clinical practice in medical decision support systems. The most common evaluation measure is accuracy that is the percentage of correctly predicted samples. However, in case of prediction on imbalanced datasets, accuracy is not a sufficient measure for evaluating classifier's performance. For example if the majority class in a dataset constitute 85% of total samples, predicting all the samples as belonging to majority class inherently yields an accuracy of 85%. Although such an accuracy level seems high, the predictor system does not provide any information about the minority class. Therefore, additional performance metrics are required to evaluate predictions for each class separately.

4.8.1. Performance Measures

In medical machine learning applications, sensitivity and specificity measures are also widely used besides the common accuracy measure [84–86]. Formal definitions for these performance criteria are given in Equations 4.10, 4.11 and 4.12 respectively and they are derived from the confusion matrix given in Table 4.2.

$$Accuracy = \frac{(TP + TN)}{(TP + FN + TN + FP)} \tag{4.10}$$

- *Sensitivity* is a measure of accuracy for correctly detecting the embryos that will implant and is equal to the ratio of number of true positives (TP) to the sum of

Table 4.2. Confusion matrix *TP:True Positives, FN:False Negatives, FP:False Positives, TN:True Negatives*

	Predicted	
Actual Case	Implanted	Not-implanted
Implanted	TP	FN
Not-implanted	FP	TN

true positives and false negatives (FN).

$$Sensitivity = \frac{(TP)}{(TP + FN)} \tag{4.11}$$

- *Specificity* represents the number of true negatives (TN) over the sum of true negative and false positives (FP) and means correctly detecting the embryos that will not implant.

$$Specificity = \frac{(TN)}{(TN + FP)} \tag{4.12}$$

Sensitivity and TPR have been used interchangeably in this thesis report as well as false alarm rate and FPR.

4.8.2. ROC Analysis

In the machine learning community, after realization of the weakness of simple accuracy rate as a performance measure, the use of ROC curves [87] have gained an increasing attention.

The ROC curve plots the sensitivity (i.e. a measure of accuracy for correctly detecting the embryos that will implant) versus (1-specificity), (i.e. erroneous posi-

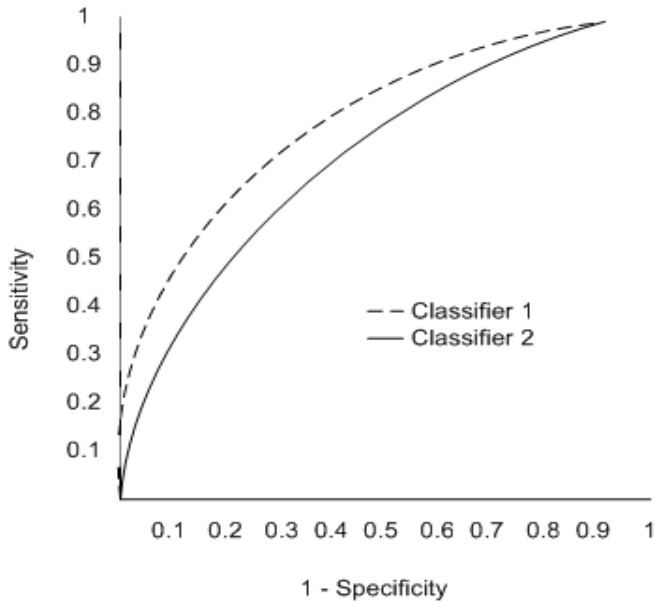


Figure 4.1. An artificial ROC curve illustrating two classifiers: Classifier 1 has larger AUC than classifier 2

tive implantation prediction) by adjusting the decision threshold of classification and enables comparison of classifiers using a single performance measure that is the area under the curve (AUC) [88].

Higher sensitivity and lower false alarm (1-specificity) rates were targeted in embryo implantation prediction; therefore the classifier with the largest AUC dominates the others. Figure 4.1 shows an example ROC curve where classifier 1 performs better than classifier 2 in terms of AUC.

It has been shown that, the AUC represents the most informative and objective performance measure within a benchmarking context [68] especially in case of imbalanced class distributions [37]. The dataset used in this research represents an imbalanced nature consisting of 89% not-implanted and 11% implanted embryos. Hence, classifier comparison and feature subset selection have been performed according to AUC measure.

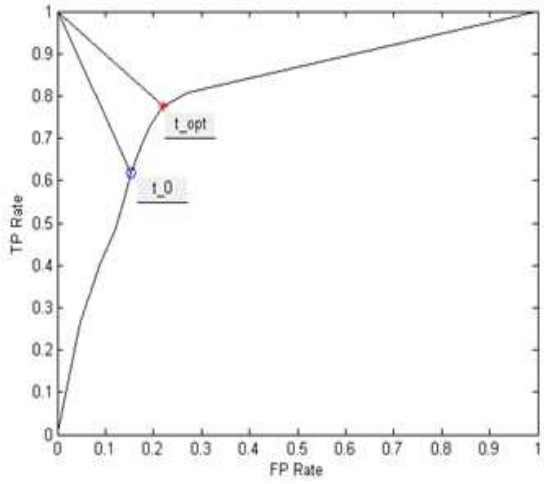


Figure 4.2. An ROC curve illustrating the effect of threshold optimization: Default threshold (t_0) and optimum threshold (t_{opt})

4.9. Post-processing

4.9.1. Decision Threshold Optimization

Nave Bayes classifier computes the class posterior probabilities, $P(C_i|x)$ of input embryo data (x) for both negative and positive implantation classes. In case of binary classification, the default decision threshold was 0.5 and the embryo was decided to belong to the class with the highest posterior probability.

The TPR and FPR have been calculated for a single threshold (default 0.5) that maps to a single point on the ROC curve. However, Provost clearly defined that, "when studying problems with imbalanced data, using the classifiers produced by standard machine learning algorithms without adjusting the output threshold may well be a critical mistake" [38]. Since, the datasets that we have utilized in this study represent imbalanced class distributions of positive (89%) and negative (11%) implantation classes of embryos, it is necessary to evaluate the performance of classification for different thresholds. We need to determine the optimum probability threshold considering both sensitivity and false alarm rates.

It is necessary to mention critical points on the 2D ROC curve. The lower left point (0,0) represents assigning all instances to negative class. Hence, there are no positive predictions yielding TPR and FPR to be 0. Conversely, upper right corner (1,1) indicates positive prediction for all instances. The upper left point (0,1) represents perfect classification. Therefore, the threshold value that gives the nearest point to (0,1) is accepted as the optimum decision threshold (t_{opt}) 4.2. Choosing a point on the left-hand side of the t_{opt} reduce false alarms but often have lower TP rates as well. Thresholds on the right hand-side increase both FP and TP rates. The tradeoff between TP and FP rates depends on the requirements of the specific application domains. Minimum distance optimization method assumes equal misclassification costs.

Embryo selection process is expected to produce higher sensitivity rates, because we do not want to miss the embryos that will implant. However, increasing sensitivity also increase false alarms that is incorrectly detecting not-implanted embryos. Probability of false alarms corresponds to (1 - specificity) and desired to have low values.

5. EXPERIMENTS AND RESULTS

5.1. Experiment I : Benchmarking Classifiers for Implantation Prediction (Research Question I)

The experiments presented in this section corresponds to the first research question: How can we construct an efficient embryo-based implantation prediction model?

5.1.1. Implantation Prediction as a Supervised Classification Problem

For each embryo, a data feature vector including 18 patient and embryo characteristics are labeled as either "implanted" (C_{+1}) or "not-implanted" (C_{-1}). Then, a classification algorithm is applied to learn a model from the training data vectors. The output of the classifier is a prediction on implantation outcome of embryos.

The algorithm for training and testing of classifiers on IVF dataset is given in Figure 5.1.

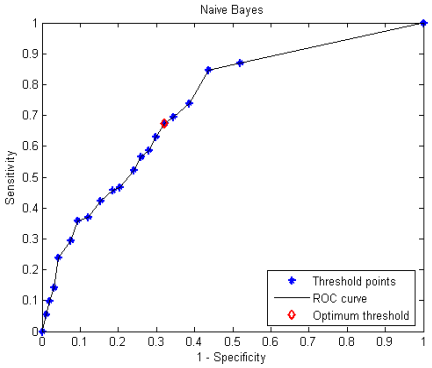
5.1.2. Results

5.1.2.1. Retrospective Analysis. All the values of TPR and FPR have been calculated by varying the decision thresholds in the range of [0:0.05:1]. The resulting set of (TPR, FPR) pairs are plotted as a 2D ROC curve that takes into account all possible solutions of the threshold variation. Among six methods, Naive Bayes and RBF were significantly better while kNN and DT were significantly worse ($P < 0.05$). The results (excluding kNN and DT) were plotted as ROC curves, which appear in Figure 5.2 demonstrating the effect of threshold optimization on the variation of TPR and FPR.

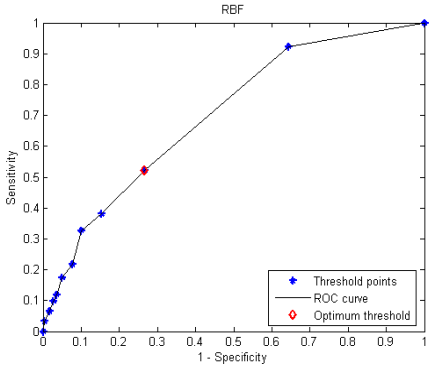
Naive Bayes classifier is used in the implantation prediction experiments due to the superior performance. Maximum sensitivity and minimum false alarm rates are desired in implantation prediction which corresponds to upper left corner (0,1) on the

1: D = [IVF Data]
2: C = [Naive Bayes, DT, kNN, SVM, MLP, RBF]
3: % generate 10 random 2/3-1/3 split of dataset
4: **for all** i in [1:1:10] **do**
5: train[i] = random 2/3 of D
6: test[i] = D - train[i]
7: **end for**
8: % pre-processing, parameter optimization and classification
9: **for all** c in C **do**
10: **for all** i in [1:1:10] **do**
11: train = train[i]
12: test = test[i]
13: $[train_p, test_p]$ = preprocess(train,test)
14: $param^*$ = arg max$_{param}$ AUC(10 fold CV on $train_p$)
15: model = learn($train_p, c, param^*$)
16: AUC[i] = classify($test_p, model$)
17: **end for**
18: output [mean(AUC), std(AUC)]
19: **end for**

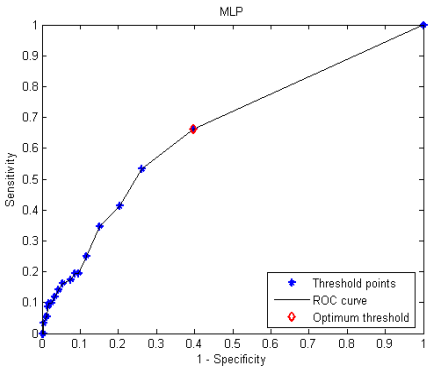
Figure 5.1. Pseudocode for evaluation of classifiers on IVF dataset



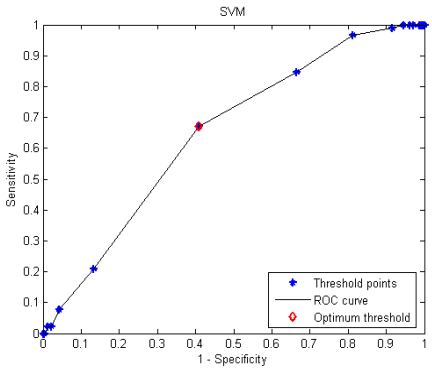
(a) ROC curve for Naive Bayes classification. AUC is 0.739 ± 0.036



(b) ROC curve for RBF classification. AUC is 0.712 ± 0.036



(c) ROC curve for MLP classification. AUC is0.675 ± 0.039



(d) ROC curve for SVM classification. AUC is 0.657 ± 0.020

Figure 5.2. ROC analysis representation for IVF dataset

ROC curve. The threshold value that maps to the point nearest the (0,1) point was estimated as 0.2 and decided to be the optimum threshold for classification.

The selected classifier specific model parameters were:

- $k = 9$ for k-NN method,
- cost, $C = 30$ and $\gamma = 10^{-4}$ in Gaussian Kernel for SVM method,
- number of clusters $B = 2$ and minimum standard deviation of clusters $w = 0.1$ for RBF method, and
- 1 hidden layer, 10 hidden units, 20 epochs with a learning rate, $\eta = 0.3$ and momentum, $\mu = 0.2$ for MLP method.

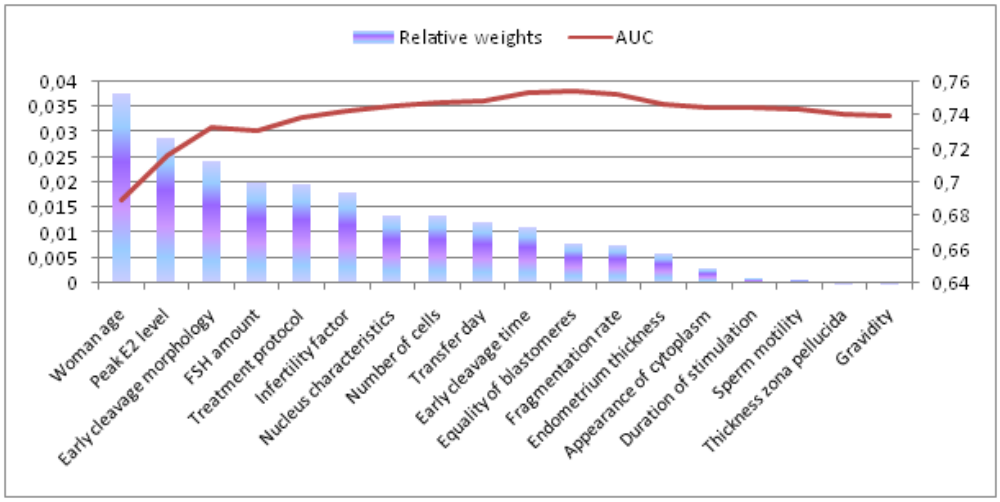


Figure 5.3. Relative weights of dataset features in decreasing order (bar graph associated with left axis) and variation of AUC depending on feature subset selection (line graph associated with right axis)

5.1.2.2. Weighted Features and Reduced Subset. Relative Information Gain weights of dataset features are given in Figure 5.3. According to feature weights, duration of stimulation, sperm motility, thickness of the zona pellucida and gravidity variables have very little predictor value on the implantation outcome. The predictive value of each data feature has been investigated using forward subset selection. Figure 4 represents the variation of prediction performance in terms of AUC displaying age of the woman as the most predictor variable.

The results show that prediction with the subset including the first 11 features produce the highest mean AUC score; 0.754, while the mean AUC with the complete feature set was 0.739. Therefore, remaining features have been discarded from the dataset in the rest of the experiments. Using these 11 features and with a decision threshold of 0.2, Nave Bayes classifier predicted the outcomes of individual embryos at 80.4% accuracy, 63.7% sensitivity and 17.6% false alarm rate. The model also predicts occurrence of triple pregnancies at 62.9% (39/62) sensitivity level and twin pregnancies at 65.6% (21/32) sensitivity level. Hence, multiple pregnancies can also be avoided by using our proposed model.

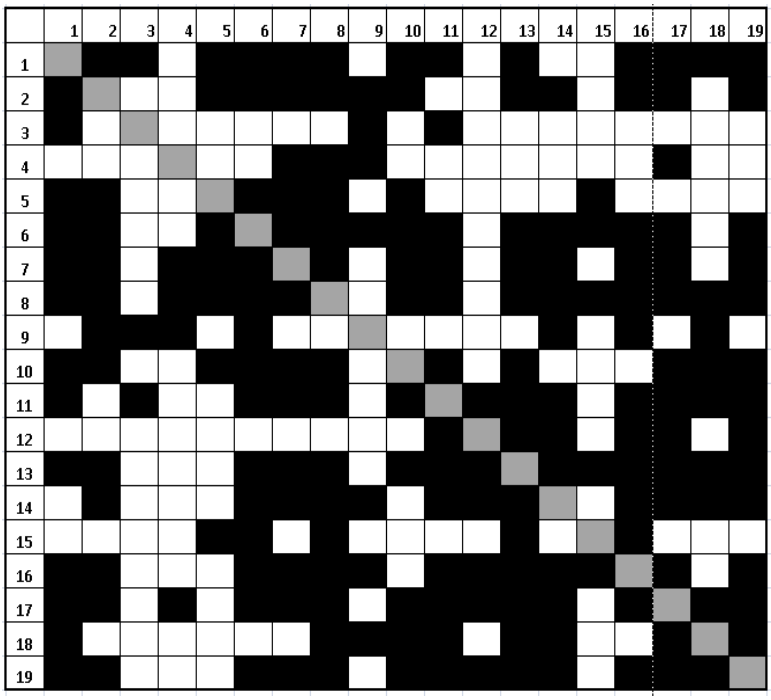


Figure 5.4. Dependency pattern of features indicating statistically significant correlations as black squares

All experiments were performed in WEKA machine learning tool [89].

5.1.2.3. Correlation Matrix and PCA. We have used correlation matrix to define the dependency structure of the dataset. We transformed the categorical variables into numerical values using our proposed frequency based transformation technique as described in Section 4.6.2.2. The resulting dataset was normalized to zero mean and unit variance. The correlation coefficients are computed on the transformed dataset.

In order to test the hypothesis of no correlation, the p-values are computed by transforming the correlation to create a t statistic having n-2 degrees of freedom, where n is the number of rows of correlation matrix.

Table 5.1. Inter-feature correlation coefficients

	1	2	3	4	5	6	7	8	9	10	11	12	3	14	15	16	17	18	19
1	1.00	-0.22	0.11	-0.02	0.06	-0.31	0.30	-0.28	0.00	-0.06	-0.08	0.01	-0.08	0.01	0.03	-0.31	-0.12	0.05	-0.23
2	-0.22	1.00	-0.01	-0.03	-0.09	0.10	-0.13	0.16	0.05	0.06	0.03	0.03	0.06	0.07	-0.02	0.16	0.12	0.01	0.09
3	0.11	-0.01	1.00	0.01	0.00	0.03	-0.01	0.00	0.15	-0.01	-0.05	-0.01	0.02	0.00	-0.02	-0.02	0.01	0.03	-0.01
4	-0.02	-0.03	0.01	1.00	-0.01	0.00	-0.05	0.05	-0.06	-0.01	-0.02	0.02	-0.04	0.00	0.04	0.01	-0.07	-0.03	0.01
5	0.06	-0.09	0.00	-0.01	1.00	-0.17	0.25	-0.05	-0.01	-0.04	-0.01	0.00	-0.01	-0.02	-0.05	-0.02	-0.03	0.00	-0.03
6	-0.31	0.10	0.03	0.00	-0.17	1.00	-0.43	0.30	0.10	0.06	0.07	0.03	0.10	0.06	-0.05	0.31	0.17	0.02	0.14
7	0.30	-0.13	-0.01	-0.05	0.25	-0.43	1.00	-0.26	-0.03	-0.05	-0.09	-0.01	-0.08	-0.06	0.03	-0.24	-0.14	0.04	-0.13
8	-0.28	0.16	0.00	0.05	-0.05	0.30	-0.26	1.00	0.02	0.05	0.10	0.02	0.13	0.07	-0.05	0.40	0.16	-0.08	0.11
9	0.00	0.05	0.15	-0.06	-0.01	0.10	-0.03	0.02	1.00	-0.03	0.00	-0.04	0.03	0.05	0.00	0.08	-0.01	-0.05	0.01
10	-0.06	0.06	-0.01	-0.01	-0.04	0.06	-0.05	0.05	-0.03	1.00	0.17	-0.01	-0.09	0.01	0.00	0.01	0.31	0.08	0.06
11	-0.08	0.03	-0.05	-0.02	-0.01	0.07	-0.09	0.10	0.00	0.17	1.00	0.26	0.38	0.17	-0.03	0.19	0.39	0.06	0.11
12	0.01	0.03	-0.01	0.02	0.00	0.03	-0.01	0.02	-0.04	-0.01	0.26	1.00	0.37	0.26	-0.04	0.09	0.18	0.03	0.09
13	-0.08	0.06	0.02	-0.04	-0.01	0.10	-0.08	0.13	0.03	-0.09	0.38	0.37	1.00	0.22	-0.04	0.16	0.28	0.05	0.10
14	0.01	0.07	0.00	0.00	-0.02	0.06	-0.06	0.07	0.05	0.01	0.17	0.26	0.22	1.00	0.02	0.16	0.18	0.05	0.05
15	0.03	-0.02	-0.02	0.04	-0.05	-0.05	0.03	-0.05	0.00	0.00	-0.03	-0.04	-0.04	0.02	1.00	-0.12	-0.04	0.03	0.00
16	-0.31	0.16	-0.02	0.01	-0.02	0.31	-0.24	0.40	0.08	0.01	0.19	0.09	0.16	0.16	-0.12	1.00	0.23	-0.03	0.13
17	-0.12	0.12	0.01	-0.07	-0.03	0.17	-0.14	0.16	-0.01	0.31	0.39	0.18	0.28	0.18	-0.04	0.23	1.00	0.24	0.16
18	0.05	0.01	0.03	-0.03	0.00	0.02	0.04	-0.08	-0.05	0.08	0.06	0.03	0.05	0.05	0.03	-0.03	0.24	1.00	-0.05
19	-0.23	0.09	-0.01	0.01	-0.03	0.14	-0.13	0.11	0.01	0.06	0.11	0.09	0.10	0.05	0.00	0.13	0.16	-0.05	1.00

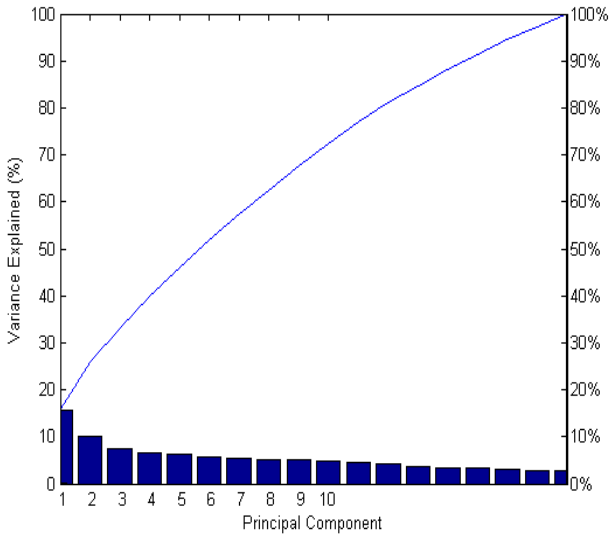


Figure 5.5. Scree Plot of PCA eigenvalues: The percent of variability explained by each principal components (vertical bar graph associated with left axis) and the cumulative percent of variability (line graph associated with right axis)

Statistically significant correlations between features are given in Figure 5.4. The 19th row and column corresponds to class variable and the feature-class correlations are also evaluated. Black squares indicate pairs with significant correlations ($p < 0.05$); the squares on the diagonal are colored gray as reference and all other squares are white. Figure 5.4 demonstrates the highly dependent structure of the IVF data.

Feature subset obtained using correlation analysis include additional two features: Fragmentation rate and appearance of cytoplasm of embryos at Day 2. The remaining 11 out of 18 features are the same in both subsets although the ranking of features according to relative predictive effects are different. Results indicate that both methods produce similar results in terms of feature selection in IVF data.

The results of feature inter-correlation analysis showed that the features are strongly dependent. Since Naive Bayes assumes independence of features, we applied PCA to extract uncorrelated features. The percent of variability explained by each principal component is shown as a scree plot in Figure 5.5.

The analysis of feature-class correlations led us to compare the predictive features selected using two different approaches: correlations analysis and Information Gain heuristics. Table 5.2 represents the list selected features in decreasing order according to estimated Information Gain values and correlation coefficients.

We have repeated the Naive Bayes classification of implantation using only those principal components which contribute more than 2% of the total variation in the transformed dataset. In that case, all the principal components retained in the dataset. The classification using 18 principal components resulted in an AUC value of 0.673 ± 0.023 where the result of classification using the 18 original input features was 0.739 ± 0.036. The classification using PCA produced significantly worse performance compared to raw IVF data.

The relatively poor performance of PCA in IVF data may also be perceived from the scree plot where the first three principal components explain only roughly one third of the total variability.

5.1.2.4. Semi-Prospective Experiments. In a semi-prospective study we asked five embryologists in Bahceci Clinic to predict the implantation outcome of embryos that were going to be transferred from May 2009 within two months. We took the majority decision of five of them as the 'expert judgement'. We also simultaneously used our model to make future predictions. After 12 weeks of embryo transfer we compared expert judgement, and model predictions with the actual outcomes. Again, for exact traceability of individual embryo implantations, we have only analyzed the cycles where all of the transferred embryos implanted or not implanted. We looked at in total of 173 embryos taken from 64 cycles including 1 single pregnancy, 6 twin pregnancies, 10 triple pregnancies and 47 negative pregnancies.

The predictions of the proposed model on 173 embryos have been presented in Table 5.3 as a confusion matrix indicating 75.7% accuracy, 55.8% sensitivity and 17.7% false alarm rate. These results are very close to retrospective analysis that validates

Table 5.2. List of selected features using Information Gain heuristic and correlation analysis

Information Gain	**Correlation Analysis**
1. Age	1. Age
2. E2 level	2. EC Morphology
3. EC Morphology	3. FSH amount
4. FSH amount	4. Treatment Protocol
5. Treatment Protocol	5. Transfer Day
6. Infertility Factor	6. Nucleus Characteristics
7. Nucleus Characteristics	7. E2 level
8. Number of cells	8. Equality of blastomeres
9. Transfer Day	9. Infertility Factor
10. EC Inspection Time	10. *Fragmentation Rate*
11. Equality of blastomeres	11. Number of cells
	12. EC Inspection Time
	13. *Appearance of cytoplasm*

Table 5.3. Confusion matrix for semi-prospective analysis

		Predicted	
Actual Case	# Embryos	Positive	Negative
Positive	43	24	19
Negative	130	23	107
Total	173	47	126

the predictive power of the proposed model.

Table 5.4. Summary of retrospective and semi-prospective experiments

Dataset	**# features**	**Threshold**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Original dataset (default threshold)	18	0.5	77.0	58.9	79.3
Reduced subset (default threshold)	11	0.5	77.2	53.7	80.1
Reduced subset (optimized threshold)	11	0.2	80.4	63.7	82.4
Semi-prospective prediction	11	0.2	75.7	55.8	82.3

All of the five embryologists had a common implantation prediction on only 28.9% of 173 embryos. This diversity indicates the influence of human bias in critical decisions that would affect the success of the IVF treatment. Our results showed that experts failed to correctly predict the implantation potential of embryos by 39.9%, however, our learning based proposed model only failed by 24.3%. Moreover, the false alarm (false positive) rate of experts is 40.8%, whereas, the false alarm rate of the proposed model is 17.7%. We would like to achieve a low false alarm rate in a prediction model since we do not want to misclassify a poor quality embryo and or poor respondent as a successful implantation. Such a misclassification has severe cost and moral implications on the patients as well as on the clinics.

The results of retrospective and semi-prospective experiments are summarized in Table 5.4.

5.1.2.5. Predictions on Random Cases. Table 5.5 represents samples of random cases in semi-prospective predictions including both the prediction of the model and expert judgement. The proposed model can predict the outcome accurately when given a good responder young patient and high quality embryos (embryo1) or poor responder older patients and low quality embryos (embryo2) as expected. These results are the predictions of the embryologists as well. However, there are some odd cases where both

experts and the proposed model failed to predict the outcome (embryos 3&4). These cases should be further investigated in order to understand the underlying reason for false predictions.

On the other hand, it is difficult to predict the outcomes of more complex cases since decision making depends on analysis of correlations between various input features. Embryos 5 and 6 are samples of such complex cases where our proposed model correctly predicts the outcome in contrast to expert judgement. These embryos represent almost similar embryo morphology with different maternal ages and infertility factors. For example, embryo 5 belonging to woman aged 33 with poor responses is classified as no-implant by the experts. However, they went ahead and transferred the embryo. On the other hand, our model predicted as an implant. The other case is embryo 6 which was transferred to a younger woman as clearly marked an implant by the experts. Our model correctly classified it as no-implant. These examples indicate that a learning based model can aid embryologists in making the right decisions in such complex cases. Such a model can learn from past experiences (i.e. thousands of embryo and patient characteristics) and makes inferences among these attributes correctly.

Table 5.5. Sample embryo feature vectors together with expert decision and predicted outcome

Embryo **ID**	Age of Woman	E2 [1]	EC Morp [1]	FSH	Treatment Protocol	Infertility Factor	Number of Nucleus	Cells	ET Day [1]	EC Time [1]	Equality of Blastomeres	Actual Outcome	Expert Decision	Predicted Outcome
1	22	2842	EY	2175	Antagonist	Male	Mono	4	3	25	Even	1	1	1
2	40	2566	2PN	4850	Antagonist	DOR	Invisible	2	3	27	Even	-1	-1	-1
3	36	968	2PN	3150	Hybrid	Unexplained	Multi	3	3	26	Uneven	1	-1	-1
4	27	3140	2PN	2650	Antagonits	Azospermia	Mono	4	3	22	Even	-1	1	
5	33	3461	2PN	2150	Long	Azospermia	Mono	4	3	27	Uneven	1	-1	1
6	28	2414	Syngami	3600	Antagonist	Endometirosis	Mono	4	3	25	Uneven	-1	1	-1

[a](E2: peak E2 level, EC Morp.: Early cleavage morphology, ET Day: Embryo transfer day, EC Time: Early cleavage inspection time (in hours)

5.2. Experiment II: Sampling vs. Threshold Optimization (Research Question 2)

The IVF dataset we analyzed contain fewer samples with positive outcomes. Any classifier built on these dataset has much more information to identify unsuccessful IVF treatments compared to successful ones. Therefore, implantation prediction is handled as a typical case of learning from imbalanced data problem. In the experiment I, we adjusted the classification threshold to avoid the learning bias to the majority class. Alternatively, the effects of sampling methods in prediction performance have been investigated in machine learning based medical decision making applications in case of imbalanced or skewed class distribution[90–92].

We analyze the effects of re-sampling the training data and decision threshold optimization on imbalanced IVF dataset using Naive Bayes classifier. We perform over- and under-sampling in different scales and examined the classification performance on the re-balanced IVF data with the default threshold of 0.5. Analysis of under-sampling experiments also leaded to define sufficient size of embryo samples for implantation prediction that would reduce the effort spent for data collection in IVF domain.We also search for the optimum classification threshold as a post-processing stage.

Re-sampling the training data can be performed in an unsupervised manner before the classification experiments. However, the optimum threshold is determined after the classification and can not be generalized as a model parameter since the decision threshold is too sensitive to the data.

5.2.1. Data and Design

Two main sampling strategies are over-sampling that replicates instances from the minority class [35] and under-sampling where some of the instances in the majority class is removed [34].

In Experiment I, we have compared various classifiers for implantation prediction

of IVF embryos and shown that Naive Bayes produce significantly better predictive performance [93]. Therefore, we apply Naive Bayes algorithm to imbalanced IVF dataset in order to investigate the effect of sampling strategies and threshold optimization.

For over sampling, we have constructed ten training sets by replicating the positive instances while keeping the number of negative instances constant. For the first over sampling, we have created one more copy of positive instances, for the second we created two copies and so on. When constructing under sampled datasets, we have included all of the positive instances and randomly selected 1/10, 2/10... of the negative instances for each fold.

The dataset includes 2275 fresh, non-donor in-vitro human embryos transferred in Day 2 or Day 3 after ICSI. The dataset used in this study represented an imbalanced nature consisting of 1944 (85.4%) negative implantation and 331 (14.6%) positive implantation outcomes (The size of the dataset changed due to missing variables in additional features such as difficulty of embryo transfer). The random two-thirds, one-third partitioning is used for training and testing.

5.2.2. Results

Table 5.6 and Table 5.7 represent the distribution of the training set and prediction results in terms of TPR and FPR for over sampling and under sampling, respectively. Results show that both TPR and FPR increase at each fold of resampling. This can be interpreted as increasing the number of positive embryo samples and reducing the number of negative embryo samples raise the number of positive predictions.

The tradeoff between the TPR and FPR can be adjusted by changing the ratio of classes. Optimum (TPR, FPR) pair can also be obtained as explained in Section 4.9.1. These corresponds to (66.5%, 33.6%) and (65.3%, 32.1%) for over sampling and under sampling, respectively.

The TPR and FPR values have also been calculated by varying the decision

Table 5.6. Distribution of class samples and prediction results after over sampling the training data

Dataset No	1	2	3	4	5	6	7	8	9	10
# of Positive Samples	218	436	654	872	1090	1308	1526	1744	1962	2180
# of Negative Samples	1295	1295	1295	1295	1295	1295	1295	1295	1295	1295
True Positive Rate	50.8	63.0	66.5	69.2	70.5	72.3	74.1	74.9	76.0	76.8
False Positive Rate	18.0	28.7	33.6	37.2	40.4	42.9	44.9	46.1	47.3	48.8

Table 5.7. Distribution of class samples and prediction results after under sampling the training data

Dataset No	1	2	3	4	5	6	7	8	9	10
# of Positive Sample	218	218	218	218	218	218	218	218	218	218
# of Negative Samples	1295	1165	1036	906	777	647	518	388	259	129
True Positive Rate	50.8	54.2	55.4	58.1	61.1	63.7	65.3	68.2	72.6	79.1
False Positive Rate	18.0	20.2	22.0	24.5	26.2	29.8	32.1	36.0	41.4	51.3

thresholds in the range of [0:0.1:1]. The resulting set of (TPR, FPR) pairs are given in Table 5.8.

The results of over-sampling, under-sampling and threshold variation have been plotted as a single 2D ROC curve (Figure 5.6). Both sampling methods and adjustment of the decision threshold produce almost the same ROC curves demonstrating the similarity of the effects of these methods on prediction performance.

Classification with the default decision threshold, i.e. 0.5, produce 50.8% TPR and 18.0% FPR, whereas with $t_{opt} = 0.3$ TPR increased to 64.4% and FPR also increased to 30.6%. Choosing a point on the left-hand side of the t_{opt} on the ROC

Table 5.8. Prediction results depending on variation of the decision threshold

Decision Threshold	0	0.1	0.2	**0.3**	0.4	0.5	0.6	0.7	0.8	0.9	1
True Positive Rate	1	77.1	69.4	**64.4**	58.4	50.8	41.3	28.0	13.1	3.6	0
False Positive Rate	1	48.2	37.6	**30.6**	23.8	18.0	13.1	8.6	4.6	0.6	0

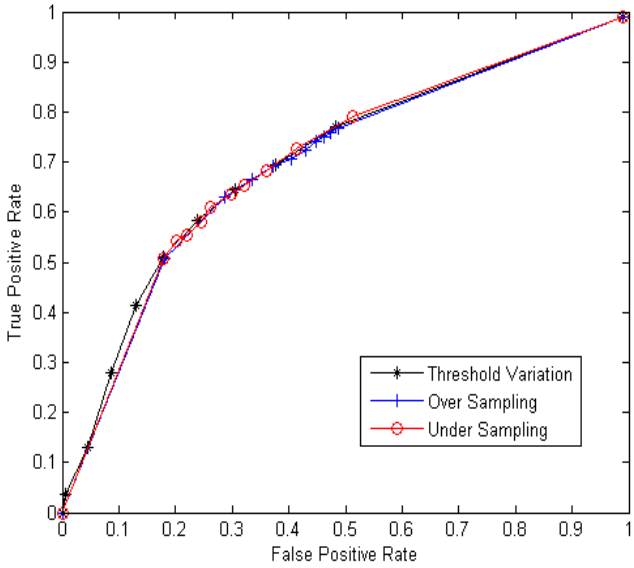


Figure 5.6. ROC curves demonstrating the effect of sampling and threshold variation of Naive Bayes based IVF implantation prediction

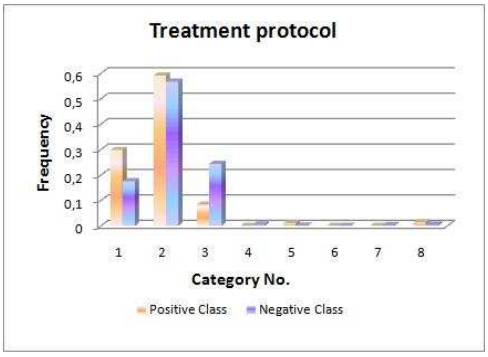
curve reduce FPR, but often have lower TPR as well. Thresholds on the right hand-side increase both TPR and FPR.

5.3. Experiment III: Transformation of Categorical Variables (Research Question II)

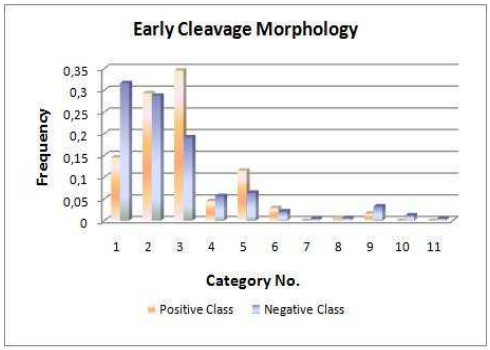
Performance of distance based classifiers, such as SVM, depends on accurate transformation of categorical variables into numeric data. In Experiment I, we used common binary encoding approach for the data type transformation. Due to the relatively poor performance of SVM classifier we examine the efficiency of binary encoding in this experiment and we propose a frequency based encoding technique for better transformation of categorical variables.

5.3.1. Data and Design

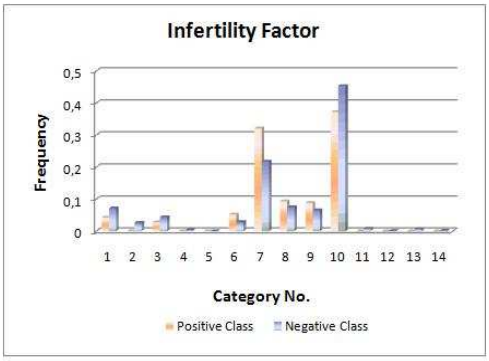
The dataset includes three categorical variables: infertility factor, treatment protocol and early cleavage morphology with 14, 8 and 11 categories, respectively. Figure



(a) Treatment Protocol



(b) Early Cleavage Morphology



(c) Infertility Factor

Figure 5.7. Distribution of categories for each categorical variable among both positive and negative implantation classes

5.7 represents the distribution of the categorical variables among both positive and negative implantation classes.

We converted categorical features into numeric values using binary encoding, expert judgement and our proposed frequency based transformation technique. After each transformation, the input data were normalized to 0 mean and standard deviation of 1. Kernel and parameter selection is performed using cross validation on the training set.

Table 5.9. Example transformation of 'treatment protocol' feature including 8 categories

Original category code	Binary encoding	Frequency based encoding	Expert judgement
1	00000001	0.123	3
2	00000010	0.024	3
3	00000100	-0.16	3
4	00001000	-0.006	2
5	00010000	0.0094	1
6	00100000	0	1
7	01000000	-0.0031	4
8	10000000	0.0086	2

5.3.2. Results

For the treatment protocols, the categories 1,2,3...8 correspond to 0.123, 0.024, -0.16...0.0086 as a result of frequency transformation as shown in Table 5.9. The frequency based encoding has the advantage of self-learning from the training set and therefore supposed to minimize the bias of transformation. This method also has the advantage of preserving the original number of features [41] since the input dimension of our dataset is increased to 42 from initial 12 features after binary encoding.

Expert judgement can be used as an alternative transformation method. This approach transforms the categories manually, making use of the domain knowledge and experience of medical specialists. The senior embryologists in Bahceci Clinic were asked to assign a numerical value to each category representing the relative predictor effect of that category on implantation outcome. They have assigned numerical values from the set of 1,2,3,4, where the greater numbers represent more predictor effect. For the treatment protocols, the manually assigned values are shown in Table 5.9. The same strategy has also been applied to early cleavage morphology and infertility factor variables. This approach may be useful for reflecting user control, however may also insert bias to the original data distribution.

The average ROC curves of SVM classification of embryos using three different

Table 5.10. Comparison of transformation methods for categorical variables

Transformation Method	**AUC**	**TP Rate (%)**	**FP Rate (%)**	**Accuracy** (%)
Binary encoding	0.676 ± 0.033	67.9 ± 4.0	37.8 ± 4.3	62.7 ± 3.7
Frequency based encoding	0.712 ± 0.032	65.6 ± 4.9	32.5 ± 7.9	67.3 ± 6.8
Expert judgement	0.696 ± 0.024	69.8 ± 8.2	36.9 ± 6.8	63.7 ± 5.5

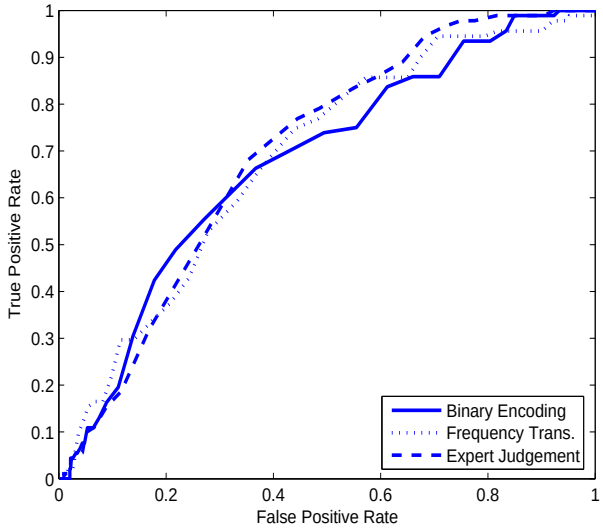


Figure 5.8. Demonstration of mean ROC curves for transformation methods

transformation schemes have been represented in Figure 5.8. For clarity, the results have also been shown in Table 5.10 in terms of AUC, TPR, FPR and accuracy. Statistical tests on the results reveal that, the proposed frequency based encoding technique significantly improves the performance of classification in AUC measure compared to binary encoding scheme (0.712±0.032 and 0.676±0.033 respectively). The values in Table 5.10 show that, proposed method increase accuracy and reduce FPR while slightly decreasing the TPR. However, these differences are not significant.

An interesting result is that, each of the three transformation methods utilized in the experiments dominates in different parts of the ROC area. This may yield to further analysis to combine the three methods for better classification performance in IVF domain.

5.4. Experiment IV: The Effect of Physicians Experience as a Human Factor (Research Question III)

We performed experiments to investigate the effect of physicians as a predictor factor other than patient and embryo characteristics in IVF treatment.

The data related to the impact of individual physicians performing embryo transfers on the PRs are conflicting [48–51]. Significant differences were observed in the studies by Hearns-Stokes et al. [49] and Karande et al. [50]. Yet, the number of embryo transfers among the physicians varied between 6 to 551 in [49] and 12 to 374 in [50]. van Weering et al. [51] analyzed the PRs from transfers performed by six physicians with similar cycle characteristics and reported that the probability of success in IVF was not dependent on the physician. The authors reported that each physician had at least 2 years of experience. However, the PRs in their study are lower than the ones in our study (19.1% to 29.0% versus 38.7% to 49.2%) and it may be more difficult to observe the differences in outcome when overall pregnancy rates are lower.

Angelini et al. citeAngelini2006 also suggested that the physician factor may be an important variable in the outcome. In contrast to studies representing a great variation of cycle distribution among the physicians, two physicians were assessed as performing 233 and 252 transfers.

5.4.1. Data and Design

Cleavage stage embryo transfer cycles that were carried out at the Bahceci IVF Centre between January 2007 and August 2009 were retrospectively analyzed. Thaw embryo transfers and PGD cycles were not included in the study.

Angelini et al. have reported that the presence of blood and mucus on the tip of the catheter was significantly more in the less experienced group of physicians although the cycle characteristics remained similar between the groups. Hence, transfers performed by this physician group yielded a lower PR. Conversely, Hearns-Stokes et

al. [49] found no evidence concerning the association of the presence of blood and the lower PRs achieved by individual physicians. The authors have also considered the type of the catheter as a confounding variable while investigating the impact of the physician on the clinical PRs and included in their study only the transfers that were performed by a single catheter type.

Because of the conflicting definitions and conclusions related to difficult transfers in the literature, we excluded the difficult transfer cycles from our study and we analyzed the embryo transfers that have been regarded as 'easy'; i.e., those which were performed by one type of catheter and those in which no blood or mucus was present on the catheter after the procedure.

Physicians performing transfers work on a rotating weekly schedule; therefore, patients were randomly assigned to each physician.

The cycle and patient demographics assessed in the experiments were the age of women, the administered FSH amount, the peak estradiol level, and the number and mean grade of the embryos transferred. The embryo quality was evaluated in relation to the number of cells, the fragmentation rate, nucleation, the equality and symmetry of blastomeres, and the appearance of the cytoplasm [94]. In short, lower scores corresponded to higher quality. All transfers were performed under ultrasound guidance to patients with full bladder using a 'soft' Wallace catheter (1816). The transfer procedure has been described in detail by Ciray et al. [47].

The clinical pregnancy rate (PR) was considered as the outcome measure. Clinical pregnancy is defined as the visualization of intrauterine gestational sac on ultrasound at 12 weeks after transfer. The differences between the PRs of individual physicians were conducted using pair wise chi-square tests. A P value of ¡.05 was considered as statistically significant. The mean age of women, mean FSH and E2 values, and the mean number and grade of embryos per cycle were compared with a one-way analysis of variance (ANOVA) followed by Tukey's multiple comparison test.

5.4.2. Results

942 clinical pregnancies were obtained from 2212 transfer cycles (42.0%). The distribution of the PRs of individual physicians' including cycle and patient demographics is given in Table 5.11. During the study period, six physicians two of whom were beginners performed the transfers (physicians 4 and 5). Four experienced physicians with similar cycle and embryo characteristics displayed similar PRs, along with one beginner (physician 4) whose transfer cycles displayed better demographics. When cycle, patient and embryo demographics were similar, the 'experienced physician' (physician 6) displayed a significantly higher PR than the 'beginner' (physician 5).

The results of the experiments show that the PR varies between the individual physicians performing the embryo transfer. A similar PR was observed in the comparison of a beginner and an experienced group when the cycles of the former were composed of high responder patients, and higher embryo qualities indicated the importance of patient characteristics and embryo quality over and above the physician factor.

If the level of experience of physicians had more impact on the outcome, the PR of beginner physicians would be significantly lower than the PRs of experienced physicians. However, physician 4 displayed similar success to experienced physicians because of the better patient and embryo characteristics. Accordingly, the patient and cycle characteristics are shown to be strong determinants in PRs.

If physician 4 is excluded, the mean age of women, mean E2 values and mean grade of embryos per cycle were similar among the patient groups of physicians. This can be interpreted as, response level of patients and quality of transfer embryos were similar among the experienced physicians (physicians 1,2,3,6) and one beginner physician (physician 5).

Table 5.11. Cycle characteristics and pregnancy rate per physician

Physician ID	N ET cycles	Age of Women [1]	FSH amount [2]	Peak E2 [3]	N Embryos	Mean N embryos per cycle [4]	Mean grade of embryos per cycle [5]	Pregnancy rate (%) [6]
1	633	34.4±5.2	3432±1678	1962±1158	1750	2.8±0.9	1.7±0.6	43.1 (273/633)
2	608	34.2±5.6	3267±1725	2026±1232	1742	2.9±0.9	1.7±0.6	45.6 (277/608)
3	233	34.2±5.6	3525±1627	2101±1277	667	2.9±1.0	1.7±0.6	41.6 (97/233)
4	317	33.4±5.5	2632±1939	2408±1298	912	2.9±0.8	1.5±0.6	46.1 (146/317)
5	173	33.9±5.7	3763±1710	1967±1113	488	2.8±0.9	1.8±0.7	38.7 (67/173)
6	248	33.7±5.7	3807±1717	2184±1339	662	2.7±0.9	1.7±0.7	49.2 (122/248)

[a]Mean age of women did not differ significantly among physicians.

[b]Mean FSH values differed significantly among physicians.

[c]Mean E2 values differed significantly, but if physician 4 excluded, E2 values did not differ.

[d]Mean N of embryos per cycle differed significantly among physicians.

[e]Mean grade of embryos per cycle differed significantly, but if physician 4 excluded, embryo grades did not differ.

[f]Pregnancy rates differed significantly only between physician 5 and physician 6 ($P = 0.02$)

The experience level of physicians 1,2,3 and 6 were also varied from 2 to 15 years. Since these physicians had similar PRs, we can conclude that the increased experience does not dramatically affect the treatment outcome after a minimum level of experience (e.g. 2 years in our clinic). The duration of sufficient training time would change according to personal skills of physicians and to the transfer procedure of the clinics.

When the response level of patients and embryo quality were similar, the PR of a beginner physician is significantly lower the highest PR. Therefore, we can conclude that if the patient and embryo characteristics are compromised, the level of physician experience may have a more determining impact on the outcome.

5.5. Experiment V: Modeling Blastocyst Development (Research Question IV)

In experiment V we aim to answer the fourth research question: How can we model the embryo growth process? And one step further, we deal with prediction of blastocyst score.

5.5.1. Prediction of Blastocyst Score as a Supervised Classification Problem

In this research, we deal with the problem of prediction of IVF blastocyst score using Bayesian Networks as a supervised classification technique. Initially, we need to identify the possible predictor variables affecting the blastocyst score. These predictor variables are represented as a multivariate input feature vector in the form of $x = (x_1, ...x_n)$. In clinical routine, embryologists observe and record morphological characteristics of embryos for the first three days of embryo growth process mentioned in Figure 2.1. Recorded morphological observations and other patient and cycle characteristics such as age of women, hormone levels etc. constitute the database for the classification problem.

The critical decision regarding the extended culture of embryos until the blasto-

cyst stage should be given at day 3. To the best of our knowledge, there is no predefined rules for the decision of extended culture. Such a critical decision is based mostly on experience of embryologists and may be assumed as random from a statistical point of view.

Finally, if the embryo is cultured until day 5, a blastocyst score is assigned using Gardner's score described in Section 2.3.3. This score is transformed to class labels in classification problem where,

- *class 1* indicates a blastocyst with a Gardner's score $\geq$3AA (i.e. high quality blastocyst and positive class in binary classification)
- *class 2* indicates a blastocyst with a Gardner's score <3AA (i.e. low quality blastocyst and negative class in binary classification).

For each instance in the dataset, Bayesian Network classifier computes $P(C_i|x_1, ...x_n)$ for both classes and assigns the instance to the class with the highest a posteriori probability.

We have constructed a dataset of 7735 blastocysts including morphological observations at day 1, day 2, day 3 and day 5 after ICSI. Initially, we have constructed the Bayesian Network topology based on Information Gain feature ranking between the input features and the blastocyst score. We have observed that, there are very few samples for some feature vectors in the conditional probability tables yielding causal insufficiency in the training phase. Therefore, we have applied a weighted nearest neighbor based approach to handle the problem of poorly learned conditional probabilities associated to the decision node. The experimental results showed that, the accuracy is increased and the false positive rate is reduced significantly while the true positive rate remained similar compared to initial Bayesian Network in prediction of blastocysts score.

In order to test the generalization ability of the proposed model, we have repeated the experiments on common datasets from UCI ML repository. The results

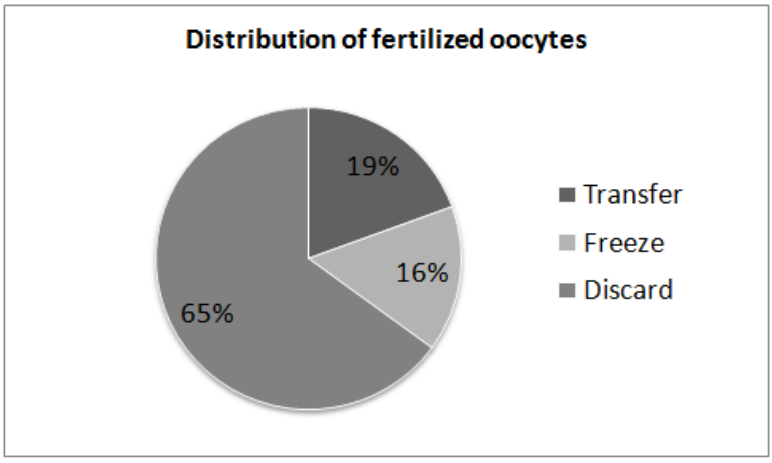


Figure 5.9. Distribution of transferred, frozen and discarded embryos in IVF cycles

showed that, the proposed method do not improve the classification performance when standard frequency estimate method already learns the conditional probabilities sufficiently. On the other hand, if insufficient frequency estimates are observed in training phase, adjusting conditional probabilities may enhance the prediction results.

5.5.2. Data and Design

There are two main advantages of Bayesian network in modeling IVF embryo growth: first, a Bayesian network can be used to learn cause-effect relationships, and hence can be used to gain understanding about the problem domain and second, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data.

We have analyzed the data of IVF cycles performed in Bahceci IVF Center from January 2007 to November 2009. Raw dataset includes a total of 81371 oocytes. Among 62800 fertilized oocytes, 12185 embryos have been transferred and 9858 embryos have been freezed (Figure 5.9). Remaining 40757 embryos have been discarded due to developmental failure that constitute 64.9% of the fertilized oocytes. This rate can be reduced by using accurate prediction models supporting the decision about extended culture of embryos although the degeneration of the embryos can not be prevented totally.

A total of 9043 embryos have been cultured until the blastocyst stage. We have eliminated the records including missing values. Finally, a total of 7735 blastocysts have been analyzed where 1779 blastocysts have been developed with a Gardners score $\geq$3AA (23.0%).

We have included the available features based on the literature and the expert judgement. The list of features is given in Table 5.12:

5.5.3. Results

5.5.3.1. Initial Bayesian Network based on Expert Judgement. In order to evaluate the performance of the Bayesian network in IVF domain, initially we have constructed the network manually based on domain knowledge to predict blastocyst score of embryos depending on morphological observations at day 3. Hence, we have a class variable as the root node that is the blastocyst score at day 5. The number of cells, nucleus characteristics and equality of blastomeres at day 3 considered as parents of root node.

The blastocyst score has originally 7 categories described in Section 2.3.3. However, the distribution of these categories is very unfair where some of the categories include only a few samples. In addition, it is more crucial to predict if blastocyst will develop (scores 2,3,4,5,6) and or will fail to develop (categories CM and 1). Therefore, in the initial experiments, the prediction of blastocyst morphology is reduced to binary classification problem.

Data is pre-processed using Matlab; visual network construction, learning the conditional probability tables from the training data and predictions on the test data have been performed using Netica software [95].

Initially, three networks have been constructed manually:

- *Network 1:* day 5 blastocyst score depends only on day 3 morphological variables.

Table 5.12. Selected dataset features for each blastocyst feature vector

Dataset Features	**Value**
Patient and Cycle Characteristics	
Woman age	Continuous
Gravidity	Primary, Secondary
Infertility factor	DOR, Endometriozis, PCOS - HPRL, Vaginismus, Hipo-Hipo, Uterine, Tubal, Azoospermi, OAT, SSS, Combined, Unexplained
Treatment protocol	Long, Low-Long, Antagonist, Natural, Femera
Duration of stimulation	Continuous
Follicular stimulating hormone dosage	Continuous
Peak Estradiol level	Continuous
Endometrium thickness	Continuous
Sperm quality	Motile, Immotile
Embryo Related Data	
Early cleavage morphology	1, 2, 3A, 3B, 3C, 3D, FRAG, 3H, BOL
Early cleavage time	Continuous
Number of cells at day 2	NC, 2, 3, ... 10, COMP, PCOMP
Nucleus characteristics at day 2	Mono, Nomono, Multinucleus, Binuclues
Fragmentation at day 3	0, 0-20%, 20-50%, ¿50%
Blastomeres at day 2	Even, Uneven
Appearance of cytoplasm at day 2	Clear, Intermediate, Granular
Number of cells at day 3	NC, 2, 3, ... 10, ≥11, COMP, PCOMP
Nucleus characteristics at day 3	Mono, Nomono, Multinucleus, Binuclues
Fragmentation at day 3	0, 0-20%, 20-50%, ¿50%
Blastomeres at day 3	Even, Uneven
Appearance of cytoplasm at day 3	Clear, Intermediate, Granular

Table 5.13. Comparison of prediction performance using different network structures

Network	**Accuracy (%)**	**TP Rate (%)**	**FP Rate (%)**
Network 1	64.9 ± 2.0	63.5 ± 5.6	33.8 ± 4.9
Network 2	58.2 ± 2.6	35.7 ± 4.2	22.6 ± 5.1
Network 3	62.7 ± 2.3	52.5 ± 4.9	29.2 ± 3.5
Network 4	65.2 ± 2.8	64.9 ± 6.7	36.1 ± 5.3

The resulting network is represented in Figure 5.10.;

- *Network 2:* includes additional links between patient characteristics and day 5 blastocyst score;
- *Network 3:* day 1, day 2 and day 3 observations are all connected to day 5 score.

The results of prediction over 10 fold cross validation is given in Table 5.13 in terms of accuracy, true positive rate (sensitivity) and false positive rate.

Paired t-tests indicate that the networks produce significantly different results in terms of accuracy, TP rate and FP rate. Network 1 performs better in predicting adequate blastocyst development while Network 2 with patient variables reduce the false positive predictions. These results reveal that, two networks perform better in different parts of the data. As the next step, different network structures may be learned from data or by a combination of prior knowledge and learning from data. The preliminary results can be enhanced by altering the network topology.

The initial networks based on prior domain knowledge represented a fairly low prediction of blastocyst development. However, the results are promising in the sense of embryo based prediction which is a more challenging issue compared to cycle based prediction.

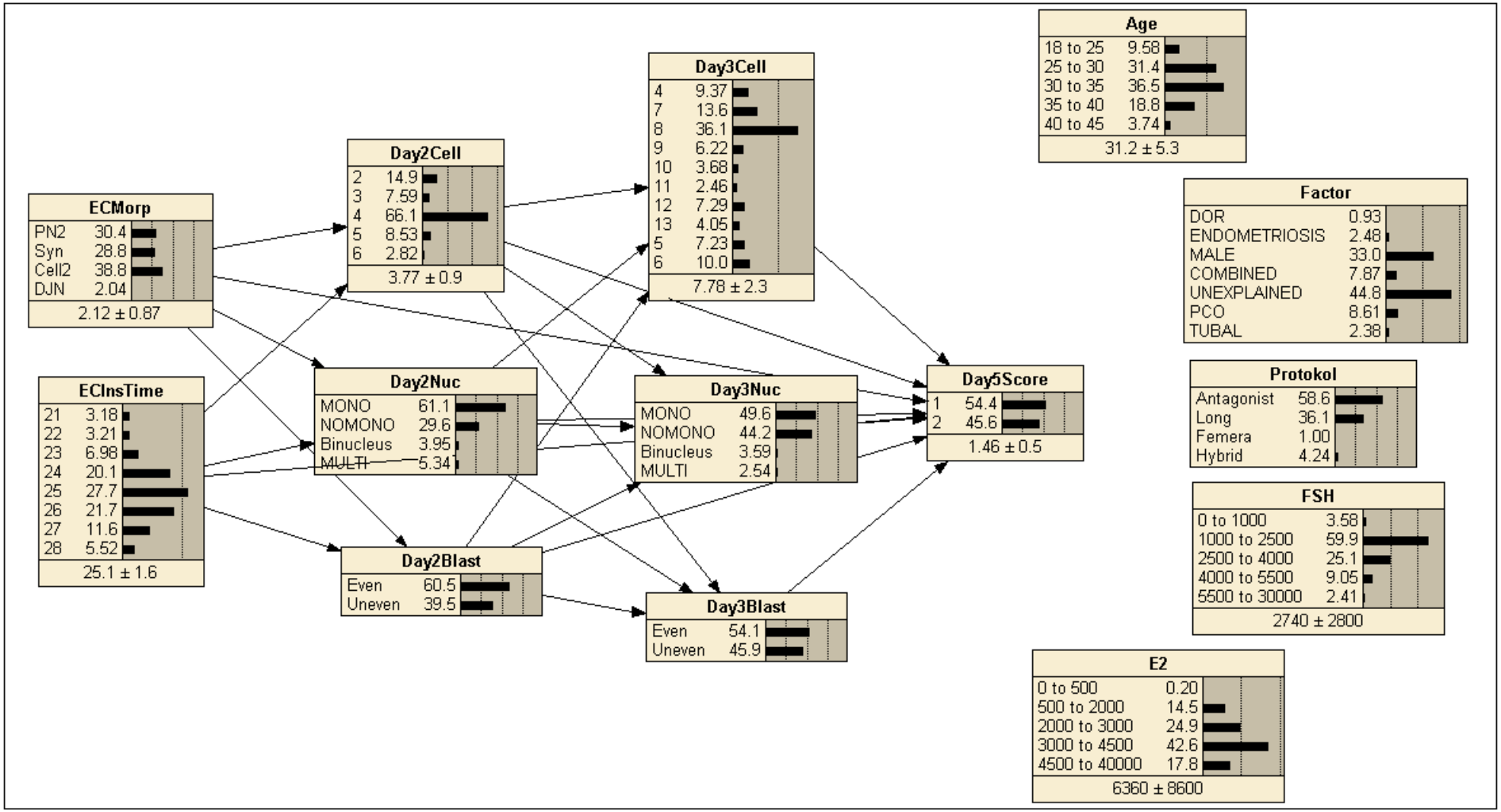


Figure 5.10. A simple embryo growth network based on correlation analysis of features

5.5.3.2. Structure Learning Based on Correlation Analysis. We have analyzed the dependency of available input features as an initial step for learning the structure of the Bayesian network from data. Correlation analysis revealed that embryo morphological variables are correlated but they are substantially independent from the patient characteristics.

The resulting network is shown in Figure 5.10. In the network, the parents of Day 5 Score include the morphological variables at day 1, day 2 and day 3. Due to the significant inter-correlations of features, there are also links between the nodes representing the daily morphological observations.

The accuracy, TPR and FPR of predictions of Day 5 score using this network (Network 4) is presented in Table 5.13. Network 4 increase accuracy and TPR rate but also increase FPR compared to Network 1. However, these differences are not significant.

The structure of the Network 4 is more complex than the other three networks because of the additional links between features. In Bayesian Networks, the size of the CPT increases depending on the complexity of the network structure. Since the parameters of the network are learnt from the data, larger CPT may result in insufficient learning and the performance of the prediction may not improve.

Therefore, we need to search for a simple network structure which encodes only necessary and sufficient relationships in IVF data and further concentrate on parameter learning.

5.5.3.3. Naive Bayesian Network and Frequency Estimate. After learning the structure of the BN, a conditional probability table (CPT) is assigned to child nodes while prior probabilities are assigned to root nodes. If n links are directed to the child node then this node has $\sum_{i=1}^{n}(N_i)$ rows in its CPT where N_i is the number of states in i_{th} root node.

In our case, we have applied Naive Bayesian network structure as a binary classifier. Bayesian networks are often used for classification problems, in which a learner attempts to construct a classifier from a given set of training instances with class labels. In learning Bayesian network classifiers, parameter learning often uses FE method, described in Section 4.5.4.2, which determines parameters by computing the appropriate frequencies from data.

In some situations there are many parents or there are many categories (and thus the conditional probability table is large) and there are few data samples to represent certain combinations of feature values. In such cases, the learning is less than optimal, and it may be necessary to find another way of estimating the probability tables.

5.6. Experiment VI: Adjusting CPT Entries for Improved Parameter Learning (Research Question V)

We consider the problem of limited samples to represent real conditional probabilities as 'partially insufficient frequency estimates'. We propose a weighted Nearest Neighbor approach to optimize the conditional probabilities to handle the insufficiency of parameter learning in Bayesian Networks.

5.6.1. Proposed Approach for Adjusting Conditional Probabilities

When the frequencies of each possible combination of feature values is computed, we can identify the samples that occur less than a predefined threshold of sample size. Then, finding the nearest neighbors of that samples constitute a cluster in the neighborhood of the infrequent sample. In this case, rather than computing the conditional probabilities for each feature vector we can compute a common conditional probability entry for the cluster of feature value combinations.

The idea behind this approach is that: Any combination of feature values may be represented insufficiently in the training data. This fact may shadow the real statistical properties of the nodes in the Bayesian Network. By clustering the less frequent

samples up to a certain level, it may be possible to obtain more accurate conditional probabilities. However, it is crucial to avoid the uniformity of conditional probabilities that would lead to information loss. Therefore, there are two critical hyper-parameters in the proposed approach:

Threshold 1: that represents the level of insufficiency in terms of frequency of feature vectors, and
Threshold 2: that represents the sufficient number of samples in the neighborhood of less frequent samples.

The thresholds should be determined in training phase using a grid search method that uses a pre-defined set of values for each threshold parameter. The search space depend on the entries in the conditional probability tables.

Nearest Neighbor approaches are generally used for classification tasks where each time a new instance needs to be classified, its similarity to the training instances is measured and the new instance inherits the class of its closest instance(s). When computing the distance between two instances, all the features may not have equal impact on the similarity measure. Therefore, identification of relative effects of the features on the distance can improve a nearest neighbor learning process [96, 97].

Feature ranking algorithms are used to identify the relevance of features in a dataset and can be used with many different distance measures. We use InfoGain feature weighting algorithm to rank the features of the dataset, and the ranked list of features is then used to define a feature weighting vector to be embedded in the Euclidean distance metric.

In this research, Nearest Neighbor approach is used for finding the most similar cases to samples which were represented less frequently in the training dataset. The weighted Euclidean distance between the instances x_i and x_j, $d_w(i, j)$ is:

$$d_w(i,j) = sqrt(\sum_{k=1}^{n}(1/w_k) * (x_{jk} - x_{ik})^2) \tag{5.1}$$

where, n is the number of features and w_k is the pre-evaluated InfoGain ranking of the k_{th} feature.

When the cluster of nearest neighbors including sufficient size of samples is obtained, the conditional probabilities are computed that average the probabilities of the samples in the cluster.

The pseudocode given in Figure 5.11 outlines the structure learning strategy that we used in network construction and our proposed approach for the parameter learning.

5.6.2. Results

5.6.2.1. Tests on IVF Dataset. Data is pre-processed using Matlab; visual network construction, learning the conditional probability tables from the training data and predictions on the test data have been performed using Netica software. The weighted nearest neighbor based post-processing of the conditional probabilities is implemented using Matlab.

Initially we have applied InfoGain feature ranking algorithm to define the network structure 5.12.

According to the structure in Figure 5.13, blastocyst score has a CPT while all other nodes have prior probabilities. There are five links directed to blastocyst score corresponding to 4*5*4*4*10 = 3200 rows in the CPT. Since there are $\sim$ 7000 samples in the trainset, certain rows have not been observed or have been occurred infrequently in the training set.

```
F = [Set of input features]
C = class variable
%Subset selection for Naive Bayesian network structure.
S = ∅
for all f in F do
    compute IG(f) = InfoGain(f, C)
    if IG(f) ≥ μ_IG(F) then
        S = S ∪ f
    end if
end for
%Frequency estimates n(Π_C = u⃗) and adjusted frequency estimates n̂(Π_C = u⃗)
%t_u upper bound for insufficient frequency and t_l lower bound for sufficient number
of data points in clustered neighborhood
for all u⃗ in S do
    if n(Π_C = u⃗) < t_u then
        n̂(Π_C = u⃗) = n(Π_C = u⃗)
        while n̂(Π_C = u⃗) < t_l do
            n̂(Π_C = u⃗) = n̂(Π_C = u⃗) + n(WeightedNearestNeighbors(u⃗))
        end while
    end if
end for
```

Figure 5.11. Pseudocode for adjusted CPT entries

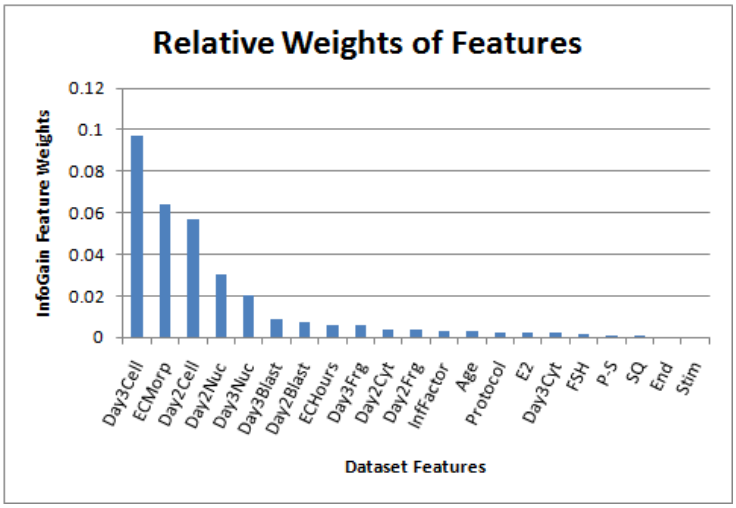


Figure 5.12. Relative Information Gain weights of features in predicting blastocyst score

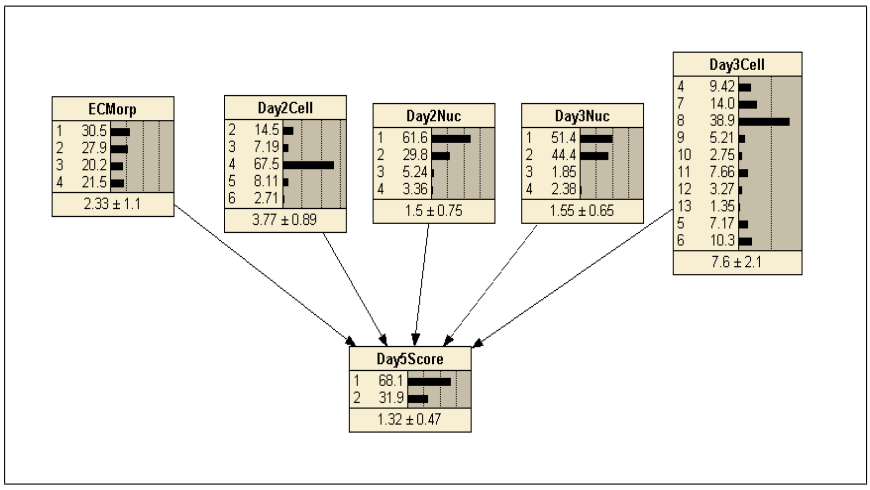


Figure 5.13. The initial Naive Bayesian network

Reducing the number of categories for some nodes may reduce the number of rows in the CPT thus may reduce the number of infrequent samples. In the initial network, number of cells at day 2 and day 3 have 5 and 10 categories, respectively. By using Netica'a auto discretization method, the numbers of categories have been reduced to 3 and 5 resulting in 4*3*4*4*5 = 960 rows in the CPT. The reduced network structure is given in Figure 5.14 where class 2 represents the high-quality blastocysts.

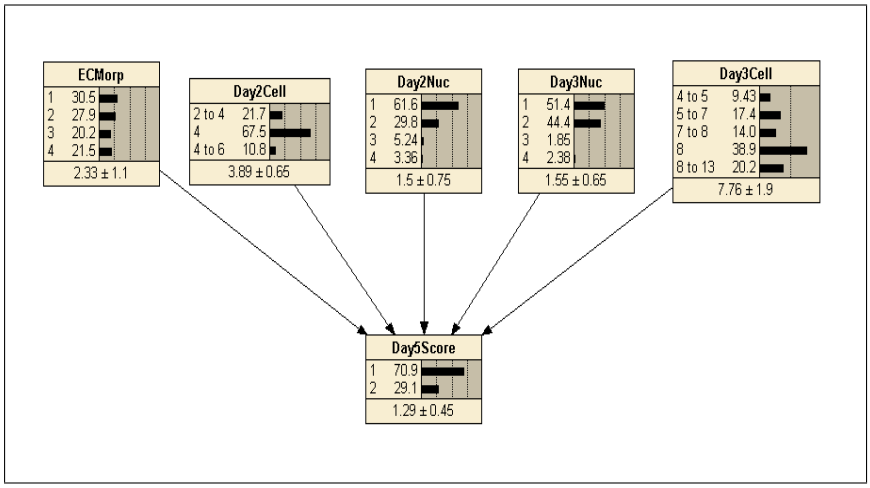
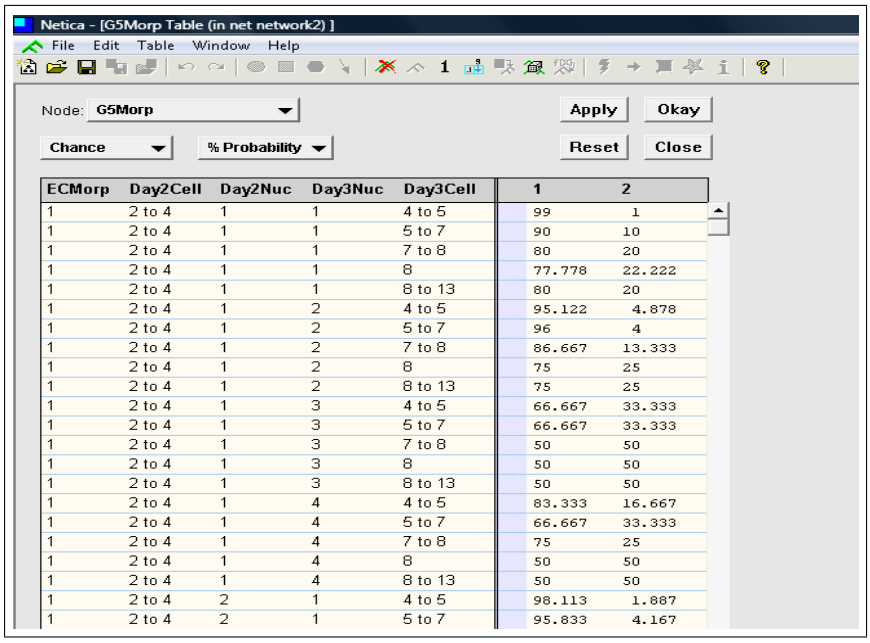


Figure 5.14. Network with reduced categories



ECMorp	Day2Cell	Day2Nuc	Day3Nuc	Day3Cell	1	2
1	2 to 4	1	1	4 to 5	99	1
1	2 to 4	1	1	5 to 7	90	10
1	2 to 4	1	1	7 to 8	80	20
1	2 to 4	1	1	8	77.778	22.222
1	2 to 4	1	1	8 to 13	80	20
1	2 to 4	1	2	4 to 5	95.122	4.878
1	2 to 4	1	2	5 to 7	96	4
1	2 to 4	1	2	7 to 8	86.667	13.333
1	2 to 4	1	2	8	75	25
1	2 to 4	1	2	8 to 13	75	25
1	2 to 4	1	3	4 to 5	66.667	33.333
1	2 to 4	1	3	5 to 7	66.667	33.333
1	2 to 4	1	3	7 to 8	50	50
1	2 to 4	1	3	8	50	50
1	2 to 4	1	3	8 to 13	50	50
1	2 to 4	1	4	4 to 5	83.333	16.667
1	2 to 4	1	4	5 to 7	66.667	33.333
1	2 to 4	1	4	7 to 8	75	25
1	2 to 4	1	4	8	50	50
1	2 to 4	1	4	8 to 13	50	50
1	2 to 4	2	1	4 to 5	98.113	1.887
1	2 to 4	2	1	5 to 7	95.833	4.167

Figure 5.15. Conditional probability table (CPT) for the blastocyst score node

Figure 5.15 shows a part of the CPT for blastocyst score node obtained from

Table 5.14. Initial probabilities in the CPT and the updated probabilities

ECMorp	D2Cell	D2Nuc	D3Nuc	D3Cell	Freq.	C1	C2	Prob.	UProb
1	1	1	1	1	98	98	0	99	99
1	1	1	1	2	8	8	0	90	92.1
1	1	1	1	3	3	3	0	80	75.9
1	1	1	1	4	7	6	1	77.778	73.8
1	1	1	1	5	3	3	0	80	70.5
1	1	1	2	1	39	38	1	95.1	94.7
1	1	1	2	2	48	47	1	96	92.4
1	1	1	2	3	13	12	1	86.6	87.0
1	1	1	2	4	2	2	0	75	79.2
1	1	1	2	5	2	2	0	75	77.4
1	1	1	3	1	1	1	0	66.6	91.9
1	1	1	3	2	1	1	0	66.6	89.0
1	1	1	3	3	0	0	0	50	86.6
1	1	1	3	4	0	0	0	50	78.9
1	1	1	3	5	0	0	0	50	76.7
1	1	1	4	1	4	4	0	83.333	79.8
1	1	1	4	2	1	1	0	66.667	79.6
1	1	1	4	3	2	2	0	75	77.7
1	1	1	4	4	0	0	0	50	76.5
1	1	1	4	5	0	0	0	50	75.8
1	1	2	1	1	51	51	0	98.113	98.1
1	1	2	1	2	46	45	1	95.833	92.4

Netica software. The probability percentages represented as 50%-50% accounts for missing observations. Hence, the CPT present the problem of insufficient statistics.

In the experiments, the CPT entry of the feature vectors that has less than 50 samples (threshold1) in the trainset have been accepted as insufficient statistics. The proposed nearest neighbor based approach has been applied to cluster the insufficiently represented CPT feature vectors to constitute a cluster of at least 200 samples (threshold2) in the trainset.

Resulting probabilities are shown in Table 5.14.

Table 5.15. Comparison of the initial network (Network1) and the network with updated CPT (Network2)

Network	**Accuracy (%)**	**TP Rate (%)**	**FP Rate (%)**
Network 1	69.1 ± 2.9	59.4 ± 7.5	29.4 ± 6.6
Network 2	72.6 ± 1.7	58.7 ± 4.8	22.7 ± 1.4

The results of prediction over stratified 10 fold cross validation is given in Table 5.15 in terms of accuracy, true positive rate (sensitivity) and false positive rate. Since the dataset represents an imbalanced distribution of the two classes of blastocysts, the decision threshold is optimized to handle the imbalance problem and decided as 0.7 mapping to the point closest to the upper left corner.

Paired t-tests indicate that the networks produce significantly different results in terms of accuracy and FP rate ($p < 0.05$). Network 2 with updated CPT reduce the false positive predictions as required in clinical procedure that would reduce the number of degenerated embryos at blastocyst stage.

5.6.2.2. Tests on UCI Datasets. Since there are no publicly available IVF datasets, we have repeated the experiments on 7 benchmark datasets from UCI ML Repository [98] to test if our model can be generalized or not.

Datasets represent a variety of data characteristics related to number of instances, number of features, number of classes, data types (continuous, discrete or mixed) and existence of missing values. Continuous variables have been discretized using the unsupervised 5-bin discretization method. Instances including missing variables have been excluded from the analysis. The multi-class datasets have been transformed to binary case by taking the two largest classes.

Information Gain feature ranking is used on the input features for each dataset and the features with the weights above the mean weight have been selected to construct the Naive Bayesian network structure.

Table 5.16. Comparison of the FE and proposed method

	FE			Proposed Method		
Dataset	AUC	TPR	FPR	AUC	TPR	FPR
Mammograph	0.82±0.04	86.5±5.4	23.3±7.6	0.83±0.04	84.5±7.1	23.8±8.2
Contraceptive	0.56±0.04	58.8±7.7	49.1±11.8	0.58±0.04	56.5±11.7	43.1±13.9
SpectHeart	0.75±0.03	68.8±3.3	19.5±6.2	0.80±0.02	77.9±3.4	24.0±4.8
Car	0.93±0.06	87.4±9.3	6.4±10.7	0.94±0.06	82.9±15.6	1.3±4.3
Voting	0.94±0.06	93.6±8.6	6.2±5.8	0.94±0.06	83.6±15.9	9.3±11.4
AustralianCredit	0.90±0.04	87.2±6.4	12.6±5.4	0.89±0.04	84.6±6.5	20.2±9.6
PimaDiabetes	0.79±0.04	73.0±7.6	26.8±4.7	0.77±0.05	69.9±7.8	26.2±4.5

The results are shown in Table 5.16.

Significance tests on the results reveal that, proposed method does not change any of the performance measures in the selected 7 datasets.

The Car, Voting and AustralianCredit datasets already represent perfect discrimination ($0.9<AUC<1$). We can conclude that, the problems that we have encountered in our dataset do not exist in these datasets. Therefore, the proposed method do not change the classification performance for these three datasets.

The Pima, Mammograph and Contraceptive datasets are large enough and the distribution of training instances over data points is fair that overcomes the problem of infrequent conditional probability entries.

The significant improvement is observed in SpectHeart dataset since the dataset characteristics and problems in frequency estimates are similar to the IVF dataset. SpectHeart dataset include 22 binary categorical variables but only 267 instances in the dataset. The number of instances in the training set is not enough to represent all the data points sufficiently in the conditional probability table as in our dataset. The proposed method increased AUC and TPR values significantly while the FPR remained the same in the SpectHeart dataset.

We can suggest that, our proposed model works well in adjusting CPT entries if the dataset characteristics satisfy our assumptions.

5.7. Discussion

The results of the Experiment I showed that a reliable implantation prediction was possible with adequate retrospective embryo-based dataset that included sufficient number of samples to train a model, prognostic data features and powerful machine learning prediction methods. The proposed model provides an implantation probability for each embryo considering both patient and embryo characteristics. Hence, unequal implantation probabilities may be assigned to any two embryos having similar cleavage morphologies but have been transferred to different patients. To the best of our knowledge, this kind of implantation prediction is novel probabilistic model in the clinical embryology literature.

The prediction performance has been improved by applying feature subset selection and classification threshold optimization. Consequently, Nave Bayes with optimized decision threshold correctly predicted the outcome with 80.4% accuracy and 63.7% sensitivity by utilizing a reduced feature set.

On the other hand, experimental results revealed the relatively lower performance of PCA in IVF data. Shlen et al. provide a clear representation of assumptions and limitations of PCA [67] that can be associated to implantation prediction problem where PCA performs poorly. Janecek et al. also concluded that the classification performance based on PCA is highly sensitive to the type of data [99].

The main assumptions of PCA are the linear correlation between the features and univariate normality which may not hold for all the features in our dataset. In addition, PCA works on continuous data where there are a number of categorical variables representing high predictive effect in our dataset. We have transformed the categorical features into numerical values using an efficient method. However, any transformation method can not preserve the whole information content of the data.

Possible information loss during the transformation together with the limitations of PCA in our data resulted in poor performance of PCA as a feature extraction method.

The model has been validated in a semi-prospective manner and the results supported the predictive power of the proposed system. It is expected that the presented implantation prediction model will provide useful information for decision-making on the number of embryos to be transferred. In situations presenting increased pregnancy probability, when the classification system predicts high implantation capacity for more than one embryo, it may be recommended to limit the number of transfer embryos. However, it should be remembered that this model may only guide the embryologist to determine the number of embryos transferred, but as there is not any confidence level it does not guarantee their selection.

Experiments II, III and IV were designed with the aim of improving the performance of implantation prediction since each real world application of standard machine learning algorithms require careful analysis of the input data and utilized methods. Selecting the most appropriate pre-processing or post-processing tasks provides better recognition performance. This is crucial for providing reliable decision support to domain experts especially in medical decision making applications.

Most of the medical datasets represent an imbalanced distribution of positive and negative samples. We examined the effects of sampling and threshold optimization in Naive Bayes classification of imbalanced datasets and presented a comparative analysis of these methods for implantation prediction of IVF embryos.

Experiment II revealed that both over sampling the minority class, under sampling the majority class and varying the decision threshold of Naive Bayes classifier produce similar prediction performance. Therefore, we suggest that, it is not necessary to artificially re-balancing the distribution of class samples in IVF dataset. The easier and effective way is to find the optimum decision threshold that produce required TPR and FPR values depending on cost of misclassifications. However, the decision threshold can not be optimized before the classification where re-sampling methods can be

performed before conducting experiments.

Under sampling experiments also show that, a training set including 218 positive and 518 negative embryo records is sufficient to characterize the implantation outcome. This result is important in the sense of reducing the time and cost of data collection in clinical practice.

Most of the medical datasets include mixed categorical and numerical attributes. We examined the effect of categorical variables in SVM classification and presented a comparative analysis of three methods for transformation of categorical variables into numeric values in mixed IVF data. We aimed to question the efficiency of traditional binary encoding method. The results of the Experiment III have shown that classification after proposed frequency transformation significantly improved the performance of SVM based implantation prediction.

According to the results obtained in Experiment IV, we concluded that patient and cycle characteristics strong determinants in success of IVF treatment. When the patient and embryo characteristics are compromised, the level of physician experience may have a more determining impact on the outcome. The PRs of experienced physicians were similar regardless of the level of experience. This can be interpreted as each physician would display average success rate in embryo transfer after a trainee period. Therefore, we can suggest that the clinical effort should concentrate on improving the response level of patients and quality of embryos.

In Experiments V we modeled the embryo growth process using Bayesian Networks with the aim of predicting blastocyst score. The initial results were relatively lower that motivated us to analyze the data and the methods. We recognized that although we have a sufficiently large dataset the observed frequency estimates are not optimal and we proposed a nearest neighbor approach to cluster the insufficient data points in Experiment VI.

There are two hyper-parameters of the proposed model: *threshold1* indicating

the lower bound for insufficient frequencies and *threshold2* indicating the upper bound for the sufficient number of training instances in the neighborhood of the infrequently represented data points. The optimum values of these two parameters depend on the distribution of training instances in the conditional probability table and the size of the dataset. Adjustment of the thresholds is critical for the success of the proposed model.

The main assumption under our proposed model is that: Infrequent or missing data points in training set can be clustered in a neighborhood to produce a more accurate collective frequency estimate for all of the instances in the associated cluster. The proposed model works well if this assumption holds. Unless, the prediction performance of frequency estimate would not change significantly. The superior performance of the proposed method adjusting the CPT entries is validated on public SpectHeart dataset.

Presented results in this research demonstrate efficacy of the prediction in terms of AUC measure. This can be interpreted as, input features were sufficient to characterize the implantation outcome providing acceptable discrimination of embryos. On the other hand, the predictive power of the presented model may be improved by increasing the information content of the input data. For instance, clinical parameters obtained from metabolomic profiling [100, 101] and pre-implantation genetic screening [102] have also shown to be effective on implantation outcomes of embryos. Moreover, it has been shown that progression to the blastocyst stage can be successfully predicted with dynamic non-invasive imaging parameters such as the time between sequential cytokinesis and mitosis [103]. Although the authors provide an approach for early diagnosis of embryo potential, we can foresee that the features obtained from time-lapse image analysis may also be used as prognostic factors on the implantation outcome. These studies assess embryo viability with novel prognostic markers providing more reliable embryo selection mechanisms.

Considering the implantation prediction problem, it is necessary to investigate both patient and embryo related variables complicating the data analysis procedure.

Our model successfully processes all the inputs together and provides acceptable discrimination of implantation outcome. The replication of our model including such chemical, genetic and time-dependent parameters in addition to clinical embryo and patient data is expected to result in higher accuracy rates compared to present study. Accordingly, the estimated predictor effects of embryo related parameters are expected to increase by applying more sophisticated embryo assessment techniques.

The proposed learning based model provides a corporate history for embryologists of a specific IVF clinic. This model integrates the experiences of all experts into a single mathematical tool since it learns from the entire dataset. It is not possible for any human expert to analyze thousands of embryo and patient records prior to each embryo transfer. However, our proposed model can perform predictions just in milliseconds and with higher accuracy than the expert judgement. Therefore, human bias and the time spent on data analysis can be minimized by using such an intelligent oracle.

This model may also be used as a self-improvement trainee tool for especially junior embryologists. In this study, we have conducted a semi-prospective study design in order to validate our model before a full-prospective automated prediction. We have compared the predictions of learning based model to the decision of embryologists and showed that we can obtain higher prediction performance by the automated predictor system.

It is important to note that dataset related pre-processing needs local tuning since the model learns from a specific dataset. For example, feature subset selection is a useful and necessary stage however the relative predictive effects of features depend on the IVF treatment process and the distribution of the dataset. In our dataset, the distributions of some of the attributes are much skewed and this situation results is lower predictive effects of those features. However, this may not be the case for other datasets.

5.8. Threats to Validity

In machine learning, it is crucial to deal with biases arising from sampling procedure and training-testing strategies. In the experiments, we used stratified cross validation in order to overcome the sampling bias. We have formed train/test sets to preserve the class distribution so that each fold can be considered as a replication of the experiments on a new dataset.

Another source of bias could arise from selection of machine learning methods. Among various classification algorithms, six models from important representatives of diverse algorithms (statistical classifiers, decision tree approaches, neural networks, support vector machines and nearest neighbor methods) have been used in this research. The prospective analysis with close prediction performance to retrospective experimental results supports the internal validity of the proposed model. Since the data used in this research comes from a single source, it is crucial to consider the external validity of the presented results. The experiment VI is validated on UCI datasets and however all the experiments need to be replicated on different IVF datasets that has no ties with the current IVF laboratory. Public dataset construction and data sharing has been a major research challenge in this domain.

The literature on application of machine learning methods in IVF domain presents conflicting results. Because there is no consensus on optimum set of input features, training and testing strategies and performance evaluation criteria. Due to imbalanced nature of IVF datasets, we have used ROC analysis and AUC, sensitivity and specificity measures. These measures are clear and widely accepted by researchers for imbalanced datasets. Finally, statistical validity is established by conducting t-tests.

6. CONCLUSIONS

6.1. Overall Summary

In this research, we present the machine learning approach as a solution to the problem of taking critical decisions under uncertainty in IVF process aiming to increase the success rates. First, we needed to understand the fundamentals of IVF treatment, the difficulty faced in decision making and the effects of these decisions on the pregnancy outcome. Then, we designed experiments in an iterative manner to match the clinical requirements to machine learning problems.

We have concentrated on two main prediction problems in IVF: predicting the implantation outcome of individual embryos and predicting whether an embryo at Day 3 will result in a high quality blastocyst at Day 5. The former problem is associated to the decision of number of embryos to be transferred and hence affects the number of multiple pregnancies. The accurate prediction for the latter problem can prevent waste of embryos and transfer cancelations arising from the developmental failure of embryos at the blastocyst stage.

The experimental design was mainly retrospective, looking back at IVF data that our collaborator clinic had been collecting for the past 4 years. We had the advantage of having a database including thousands of patient and embryo records which is a great opportunity in a machine learning study. On the other hand, there were missing or incorrect values in the database and we had to use pre-processing techniques prior to each experiment. Moreover, since we performed experiments on a dynamic database, we have reconstructed the dataset each time and therefore the dataset size and features changed during the research period as we conducted various experiments.

The clinical problems were formalized as supervised binary classifications in this research. From a machine learning perspective, the imbalanced class ratio of positive and negative samples entailed the investigation of prediction performance in terms

of TPR and FPR rather than single accuracy measure. Also, sensitivity (TPR) and specificity (1-FPR) are the common performance measures in the medical literature. Therefore, the performance criteria was based on the ROC analysis in our experiments.

Concerning the embryo based implantation prediction problem as the first research direction, we started with model selection. The forward feature selection is used to eliminate the redundant variables and hence to increase the prediction performance. The search for the optimum feature subset is based on Information Gain feature ranking rather than a random search. We dealt with two main problems of the dataset characteristics: the imbalanced class ratio and the mixed data type including both continuous numerical and categorical features. We have analyzed the effect of physicians' factor on the success of the treatment in order to improve the information content of the dataset.

After a comparative analysis of the diverse classifiers, we decided the Naive Bayes to be the best fitting algorithm for the implantation prediction problem. Naive Bayes classifier provided acceptable discrimination of the implantation outcome with a significantly higher prediction performance in terms of AUC measure.

We have designed two main experiments to improve the prediction performance by using methodological enhancements to standard machine learning algorithms. First, we compared the re-sampling methods and decision threshold optimization in order to handle the imbalance problem. Experiments showed that adjusting the classification threshold and re-sampling the training data provides similar results. Second, we proposed a frequency based encoding technique for transformation of categorical variables as a pre-processing stage to SVM classification. The prediction performance increased significantly compared to standard binary encoding technique.

With the aim of improving the information content of the data, the analysis of the physician factor indicated that patient and embryo characteristics are strong determinants in pregnancy rates. When these characteristics are compromised, the experience level of physicians performing embryo transfers may have a more determining impact

on the outcome.

The next research direction was to predict the developmental potential of the blastocysts by modeling the entire embryo growth process. We have used Bayesian Networks to predict the blastocyst score for three reasons: First, it was necessary to determine the cause-effect relationships between the daily morphological observations for an overall modeling. Bayesian Networks encode the statistical relationship between the variables of interest. Second, our dataset includes many categorical variables and Bayesian Networks provide efficient processing of categorical inputs directly. And third, Bayesian Networks enable visual representation of the underlying model which makes it more understandable for the clinicians.

The initial networks based on prior domain knowledge represented relatively low prediction performance. However, the results have been improved by enhanced structure and parameter learning. The dataset was large enough that we could learn the parameters of the CPT from the data. However, the distribution of the training data over the CPT entries was not fair that degraded the prediction capability of Bayesian Network. We proposed a nearest neighbor based approach to adjust the values in the CPT resulting in significantly higher prediction performance.

To conclude, this research presented the potential of machine learning algorithms in increasing the success rates in IVF treatment. As a preliminary comprehensive research, our results are promising in the sense of embryo based prediction. Based on our findings, we can advise using Naive Bayes for implantation prediction of IVF embryos and using Bayesian Networks for modeling developmental stages of IVF embryos. Both models need local tuning in each IVF clinic since we have performed experiments on a single IVF database. However, if the experiments are repeated on a database of a clinic with similar treatment procedures and dataset characteristics, we do not expect the results to change dramatically.

6.2. Theoretical and Methodological Contributions

The literature on using machine learning algorithms in IVF domain is very scarce probably due to the lack of public IVF datasets and poor predictive performance obtained in earlier studies. Especially there are very few studies concerning embryo based prediction. Therefore, the framework of this research was utilization of machine learning algorithms for embryo based predictions. This is an important contribution for both machine learning and IVF communities.

The preliminary experiments were application of well known classification procedures in implantation prediction problem. On the other hand, representation of the implantation prediction as a machine learning problem, unbiased stratified and cross validated training and testing strategy and evaluation of several classifiers in a comparative manner using ROC analysis was a novel approach to prediction problems in IVF treatment.

Further theoretical and methodological contributions of this research can be summarized as follows:

- *Handling the imbalance problem:* Imbalanced class distributions occur frequently in medical datasets where negative samples generally dominate positive ones. Re-sampling the training data is a common approach to balance the classes either by over-sampling the minority class or by under-sampling the majority class. However, both methods have some deficiencies and require additional computational effort as a pre-processing stage. Rather, we showed that simply finding the optimum classification threshold using ROC analysis produce similar results to re-sampling methods in implantation prediction problem. Under-sampling methods also figured out the minimum number of required positive and negative embryo records which is sufficient to characterize the implantation outcome. This result is important in the sense of reducing the time and cost of data collection in clinical practice.
- *Frequency based encoding for transformation of categorical variables:* Selecting

the best predictor model that efficiently handles both continuous and categorical variables is a challenge in machine learning applications. Some models are based on processing continuous numerical variables such as SVM and kNN while some others like Decision Trees and Bayesian Networks works better with categorical features. Data type transformation is a necessary pre-processing step in datasets consisting of mixed data types. Binary encoding is the most common approach for transformation of categorical variables into numeric values. However, it has some limitations and assumptions that can not be generalized. We proposed a frequency based encoding technique to efficiently handle categorical inputs in SVM classification. We showed that the proposed model significantly improves the prediction performance compared to binary encoding.

- *Bayesian Network modeling of blastocyst development:* Model selection is a critical and time-consuming step in machine learning studies. After evaluating the assumptions of the candidate algorithms in relation to the characteristics of the embryo growth, we presented a Bayesian Network approach for modeling the developmental stages of the embryos. The ultimate objective of this modeling was to predict the blastocyst score for individual embryos which is an open question in IVF literature. We have demonstrated the potential of Bayesian Networks in embryo based prediction of blastocyst development.
- *Adjusting the frequency estimates in CPT:* Frequency Estimate is a simple but efficient method for computing CPT parameters in Bayesian Networks. The accuracy of the frequency estimates depend on the size of the training data, size of the CPT table and distribution of the training samples. We proposed a nearest neighbor based method to adjust the frequency estimates in prediction of IVF blastocyst score. The superior performance of the proposed model has been validated on external data.

6.3. The Clinical Perspective

We suggest that the best way to construct an embryo-based prediction model is to investigate SET cycles for better identification of predictor effects of embryo variables. However, multiple embryo transfer was a routine procedure in Turkey until March 2010,

therefore our dataset includes too many multiple embryos transfers. At the beginning of 2010, the Ministry of Health published new regulations in IVF treatment allowing only SET for woman under 35 in the first or second IVF cycles.

The proposed learning based model provides a corporate history for embryologists of a specific IVF clinic. This model integrates the experiences of all experts into a single mathematical tool since it learns from the entire dataset. It is not possible for any human expert to analyze thousands of embryo and patient records prior to each embryo transfer. Therefore, the human bias and the time spent on data analysis can be minimized by using such an intelligent oracle. We believe that, such decision support systems will be common tools in IVF process due to difficulty in the analysis of increasing number of prognostic factors. This model may also be used as a self-improvement trainee tool for especially junior embryologists.

It is important to note that dataset related pre-processing needs local tuning since the model learns from a specific dataset. For example, feature subset selection is a useful and necessary stage however the relative predictive effects of features depend on the IVF treatment process and the distribution of the dataset features. In our dataset, the distributions of some of the attributes are much skewed and this situation results is lower predictive effects of those features. However, this may not be the case for other datasets. In addition, the physicians' impact on the success of embryo transfer should also be considered when constructing such prediction models.

Accurate prediction of multiple pregnancies is definitely an important contribution of the proposed model to the IVF domain. It is expected that the presented model will provide useful information for decision-making on the number of embryos to be transferred. In situations presenting increased pregnancy probability, when the classification system predicts high implantation capacity for more than one embryo, it may be recommended to limit the number of transfer embryos. At the far end, we aim to use our model to reduce multiple embryo transfers in case of multiple pregnancy risks. Such a model can provide a reliable eSET criterion to be applicable in clinical practice. This model can also be used to safely cancel embryo transfer process if the

implantation outcome is predicted as negative. In that case, high quality embryos may be frozen to be transferred in another cycle representing better response to treatment (hormone levels and endometrium, etc.).

6.4. Future Research Directions

The recent regulations forbidding multiple embryo transfers for most of the cycles in Turkey (and in the world as well since most countries have such regulations) will provide the desired SET datasets to be a base for further machine learning studies in implantation prediction. As one of the main future research directions, the experiments presented in this research should be repeated on a dataset including only SET cycles. This would better determine the predictor effects of the patient and embryo related variables by preventing the replication of patient characteristics for the embryos transferred in the same cycle and by removing the confusion about dependency of embryo implantations in case of multiple embryo transfers.

The metabolomics and pre-implantation genetic diagnosis (PGD) are becoming routine embryo assessment techniques as well as morphological observations. The features obtained from the metabolomics and genetic analysis would increase the information content of the input data both for prediction of implantation and blastocyst score. Therefore, the machine learning models should be re-trained using the extended datasets.

Automated tracking of cell divisions and automated extraction of the morphological features from continuous time-lapse embryo images is an important future research direction. After feature extraction, Hidden Markov Models or Sequential Monte Carlo methods can be used to model the embryo growth process. Automated tracking of cell divisions would better identify the potential high quality blastocysts where automated feature extraction would minimize the human bias in morphological observation.

We have mostly concentrated on parameter learning in Bayesian Networks. However, as the number of dataset features increase the structure learning becomes much

more complicated. Finding the optimum network structure that represents the embryo growth process can be further investigated.

Finally, all the results obtained in retrospective experiments may be validated prospectively as a future work.

REFERENCES

1. Mortimer, D. and S. T. Mortimer, *Quality and Risk Management in the IVF Laboratory*, Cambridge University Press, 2005.

2. Steptoe, P. C. and R. G. Edwards, "Birth After Re-implantation of a Human Embryo", *Lancet*, Vol. 2, p. 366, 1978.

3. Uyar, A., A. Bener, H. N. Ciray, and M. Bahceci, "Handling the Imbalance Problem of IVF Implantation Prediction", *IAENG International Journal of Computer Science*, Vol. 37, pp. 164–170, 2010.

4. Uyar, A., A. Bener, H. N. Ciray, and M. Bahceci, "A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset", *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6214–6217, 2009.

5. Uyar, A., A. Bener, H. N. Ciray, and M. Bahceci, "Physicians experience in performing embryo transfers may affect outcome", *Fertility and Sterility*, 2010.

6. Uyar, A., A. Bener, H. N. Ciray, and M. Bahceci, "Bayesian Networks for predicting IVF blastocyst development", *20th International Conference on Pattern Recognition (ICPR)*, pp. 2772–2775, 2010.

7. Van Steirteghem, A. C., Z. Nagy, and H. Joris, "Higher Fetilization and Implantation Rates After Intracytoplasmic Sperm Injection", *Human Reproduction*, Vol. 8, pp. 1061–1066, 1993.

8. Ciray, H., L. Karagenc, U. Ulug, F. Bener, and M. Bahceci, "Use of both early cleavage and day 2 mononucleation to predict embryos with high implantation potential in intracytoplasmic sperm injection cycles", *Fertility and Sterility*, Vol. 84, pp. 1411–1416, 2005.

9. Baczkowski, T., R. Kurzawa, and W. Glabowski, "Methods of Embryo Scoring in in vitro Fertilization", *Reproductive Biology*, Vol. 4, No. 1, pp. 5–22, 2004.

10. Shen, S., A. Khabani, N. Klein, and D. Battaglia, "Statistical Analysis Of Factors Affecting Fertilization Rates And Clinical Outcome Associated With Intracytoplasmic Sperm Injection", *American Society for Reproductive Medicine*, Vol. 79, No. 2, 2003.

11. Holte, J., 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 Embryo Scored and Transferred on day 2 after Oocyte Retrieval", *Human Reproduction*, Vol. 22, No. 2, pp. 548–557, 2007.

12. Kaufmann, S. J., J. L. Eastauh, S. Snowden, S. W. Smye, and V. Sharma, "The Application of Neural Networks in Predicting the Outcome of In-Vitro Fertilization", *Human Reproduction*, Vol. 12, pp. 1454–1457, 1997.

13. Trimarchi, J. R., J. Goodside, L. Passmore, T. Silberstein, L. Hamel, and L. Gonzalez, "Comparing Data Mining and Logistic Regression for Predicting IVF Outcome", *Fertility and Sterility*, Vol. 80, p. 100, 2003.

14. Jurisica, I., J. Mylopoulos, J. Glasgow, H. Shapiro, and R. F. Casper, "Case-Based Reasoning in IVF: Prediction and Knowledge Mining", *Artificial Intelligence in Medicine*, Vol. 12, pp. 1–24, 1998.

15. Jun, S., B. Choi, L. Shahine, L. Westphal, B. Behr, R. Pera, W. A. Wong, and M. Yao, "Defining Human Embryo Phenotypes by Cohort-Specific Prognostic Factors", *PLoS ONE*, Vol. 3(7), p. e2562. doi:10.1371/journal.pone.0002562, 2008.

16. Saith, R. R., 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", *Human Reproduction Update*, Vol. 4, No. 2, pp. 121–134, 1998.

17. Morales, D. A., E. Bengoetxea, B. Larranaga, M. Garcia, Y. Franco, M. Fresnada, and M. Merino, "Bayesian Classification for the Selection of In Vitro Human Embryos Using Morphological and Clinical Data", *Computer Methods and Programs in Biomedicine*, Vol. 90, pp. 104–116, 2008.

18. Gerris, J. and D. De Neubourg, "Single embryo transfer after IVF/ICSI: present possibilities and limits", *Journal of Obstetrics and Gynecology in India*, Vol. 55, pp. 26–47, 2005.

19. Martikainen, H., M. Orava, J. Lakkakorpi, and L. Tuomivaara, "Day 2 Elective Single Embryo Transfer in Clinical Practice: Better Outcome in ICSI Cycles", *Human Reproduction*, Vol. 19, pp. 1364–1366, 2004.

20. Veleva, Z., S. Vilska, C. Hydn-Granskog, A. Tiitinen, S. J. Tapanainen, and H. Martikainen, "Elective single embryo transfer in women aged 36–39 years", *Human Reproduction*, Vol. 21, pp. 2098–2102, 2006.

21. Thurin, A., J. Hausken, T. Hillensj, B. Jablonowska, A. Pinborg, A. Strandell, and C. Bergh, "Elective Single-Embryo Transfer versus Double-Embryo Transfer in in Vitro Fertilization", *The New England Journal of Medicine*, Vol. 351, pp. 2392–402, 2004.

22. "Society for Assited Reproductive Technology", `http://www.sart.org/`, 2010.

23. "Human Fertility and Embryology Authority", `http://www.hfea.gov.uk/`, 2010.

24. Maheshwari, A., S. Griffiths, and S. Bhattacharya, "Global variations in the uptake of single embryo transfer", *Human Reproduction Update*, Vol. 17, pp. 107–120, 2011.

25. Urman, B. and K. Yakin, "New Turkish legislation on assisted reproductive techniques and centres: a step in the right direction?", *Reproductive BioMedicine Online*, 2010.

26. van Peperstraten, A., R. Hermens, W. Nelen, P. Stalmeier, G. Scheffer, R. Grol, and J. Kremer, "Perceived barriers to elective single embryo transfer among IVF professionals: a national survey", *Human Reproduction*, Vol. 23, pp. 2718–2723, 2008.

27. van Loendersloot, L., M. van Wely, J. Limpens, P. Bossuyt, S. Repping, and F. van der Veen, "Predictive factors in in vitro fertilization (IVF): a systematic

review and meta-analysis", *Human Reproduction Update*, Vol. 16, pp. 577–589, 2010.

28. Papanikolaou, E. G., E. M. Kolibianakis, H. Tournaye, C. A. Venetis, H. Fatemi, B. Tarlatzis, and P. Devroey, "Live Birth Rates after Transfer of Equal Number of Blastocysts or Cleavage-Stage Embryos in IVF. A Systematic Review and Meta-Analysis", *Human Reproduction*, Vol. 23, pp. 91–99, 2008.

29. Gardner, D. K., M. Lane, J. Stevens, T. Schlenker, and W. B. Schoolcraft, "Blastocyst Score Affects Implantation and Pregnancy Outcome: Towards a Single Blastocyst Transfer", *Fertility and Sterility*, Vol. 73, pp. 1155–1158, 2000.

30. Dessolle, L., T. Frour, P. Barrire, E. Dara, R. C., M. Jean, and C. Coutant, "A Cycle-Based Model to Predict Blastocyst Transfer Cancellation", *Human Reproduction*, Vol. 25, pp. 598–604, 2010, accepted for publication.

31. Gardner, D. K., E. Surrey, D. Minjarez, A. Leitz, J. Stevens, and W. B. Schoolcraft, "Single Blastocyst Transfer: A Prospective Randomized Trial", *Fertility and Sterility*, Vol. 81, pp. 551–555, 2004.

32. Huang, K., H. Yang, I. King, and M. Lyu, "Maximizing Sensitivity in Medical Diagnosis Using Biased Minimax Probability Machine", *IEEE Transactions on Biomedical Engineering*, Vol. 53, pp. 821–831, 2006.

33. Mena, L. and J. Gonzalez, "Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic", *19th International FLAIRS Conference (FLAIRS-2006)*, Melbourne Beach, Florida, May 11-13 2006.

34. Kubat, M. and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection.", *Fourteenth International Conference on Machine Learning*, pp. 179–186, San Francisco, CA: Morgan Kaufmann., 1997.

35. Ling, C. and C. Li, "Data mining for direct marketing: Problems and solutions", *Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98)*, pp. 73–79, AAAI Press, Menlo Park, CA, 1998.

36. Chawla, N., K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Syntethic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.

37. Maloof, A. M., "Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown", *Workshop on Learning from Imbalanced Data Sets*, 2003.

38. Provost, F., "Machine Learning from Imbalanced Data Sets 101", *Working Notes AAAI00 Workshop Learning from Imbalanced Data Sets*, pp. 1–3, 2000.

39. Brouwer, R., "A Hybrid Neural Network with Fuzzy Rules for Categorical and Numeric Input", *International Journal of Intelligent Systems*, Vol. 19, pp. 979–1001, 2004.

40. Rogovschi, N., M. Lebbah, and Y. Bennani, "Probabilistic Mixed Topological Map for Categorical and Continuous Data", *Seventh International Conference on Machine Learning and Applications*, 2008.

41. Orsenigo, C. and C. Vercellis, "Predicting HIV Protease-Cleavable Peptides by Discrete Support Vector Machines", *EvoBIO*, 2007.

42. Ninomiya, T., "Clustering Observations using Fuzzy Similarities between Ordered Categorical Data", *Systems, Man and Cybernetics, IEEE Int. Conf. on*, 2005.

43. Tomas, C., K. Tikkinen, L. Tuomivaara, J. Tapanainen, and H. Martikainen, "The Degree of Difficulty of Embryo Transfer is an Independent Factor for Predicting Pregnancy", *Human Reproduction*, Vol. 17, pp. 2632–2635, 2002.

44. Schoolcraft, W. B., E. S. Surrey, and D. K. Gardner, "Embryo transfer: techniques and variables affecting success", *Fertility and Sterility*, Vol. 76, pp. 863–870, 2001.

45. van Weering, H. G., R. Schats, J. McDonnell, J. M. Vink, J. P. Vermeiden, and P. G. Hompes, "The Impact of the Embryo Transfer Catheter on the Pregnancy Rate in IVF", *Human Reproduction*, Vol. 17, pp. 666–670, 2002.

46. Goudas, V. T., D. G. Hammitt, M. A. Damario, D. R. Session, A. P. Singh, and D. A. Dumesic, "Blood on the Embryo Transfer Catheter is Associated with

Decreased Rates of Embryo Implantation and Clinical Pregnancy with the use of In Vitro Fertilization – Embryo Transfer", *Fertility and Sterility*, Vol. 70, pp. 878–882, 1998.

47. Ciray, H. N., S. Tosun, O. Hacifazlioglu, A. Mesut, and M. Bahceci, "Prolonged Duration of Transfer Does Not Affect Outcome in Cycles with Good Embryo Quality", *Fertility and Sterility*, Vol. 87, pp. 1218–1221, 2007.

48. Angelini, A., G. F. Brusco, N. Barnocchi, I. El-Danasouri, A. Pacchiarotti, and H. A. Selman, "Impact of Physician Performing Embryo Transfer on Pregnancy Rates in an Assisted Reproductive Program", *Journal of Assisted Reproduction and Genetics*, Vol. 23, pp. 329–332, 2006.

49. Hearns-Stokes, R., B. Miller, L. Scott, D. Creuss, P. Chakraborty, and J. Segars, "Pregnancy Rates after Embryo Transfer Depend on the Provider at Embryo Transfer", *Fertility and Sterility*, Vol. 74, pp. 80–86, 2000.

50. Karande, V. C., R. Morris, C. Chapman, J. Rinehart, and N. Gleicher, "Impact of the Physician Factor on Pregnancy Rates in a Large Assisted Reproductive Technology Program: Do too many Cooks Spoil the Broth?", *Fertility and Sterility*, Vol. 71, pp. 1001–1009, 1999.

51. van Weering, H. G., R. Schats, J. McDonnell, and P. G. Hompes, "Ongoing Pregnancy Rates in In Vitro Fertilization are not Dependent on Physician Performing the Embryo Transfer", *Fertility and Sterility*, Vol. 83, pp. 316–320, 2005.

52. Lucas, P., L. van der Gaag, and A. Abu-Hanna, "Bayesian networks in biomedicine and health care", *Artificial Intelligence in Medicine*, Vol. 30, pp. 201–214, 2004.

53. van der Gaag, L., S. Renooij, A. Feelders, A. de Groote, M. Eijkemans, F. Broekmans, and B. Fauser, "Aligning Bayesian Network Classifiers with Medical Contexts", Technical report, Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands, 2008.

54. Reiz, B. and L. Csat, "Tree-Like Bayesian Network Classifiers for Surgery Survival Chance Prediction", *International Journal of Computers, Communications and Control*, Vol. 3, pp. 470–474, 2008.

55. Antal, P., H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote, "Bayesian networks in ovarian cancer diagnosis: potentials and limitations", *Computer-Based Medical Systems, 2000. CBMS 2000. Proceedings. 13th IEEE Symposium on*, pp. 103 –108, 2000.

56. Hunt, M., B. von Konsky, S. Venkatesh, and P. Petros, "Bayesian Networks and Decision Trees in the Diagnosis of Female Urinary Incontinence", *22nd Annual EMBS International Conference*, 2000.

57. Visscher, S., P. Lucas, C. Schurink, and M. Bonten, "Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions", *Artificial Intelligence in Medicine*, Vol. 46, pp. 251–266, 2009.

58. Cheng, J., R. Greiner, J. Kelly, D. Bell, and W. Liu, "Learning Bayesian Networks from Data: An Information-Theory Based Approach", *Artificial Intelligence*, Vol. 137, pp. 43–90, 2002.

59. "Clinical epidemiology & evidence-based medicine glossary: clinical study design and methods terminology", `http://www.vetmed.wsu.edu/courses-jmgay/glossclinstudy.htm`, 2010.

60. Tiersma, E., M. van der Lee, A. Peters, A. Visser, G. Fleuren, B. Garssen, K. van Leeuwen, S. le Cessie, and K. Goodkine, "Psychosocial factors and the grade of cervical intra-epithelial neoplasia: a semi-prospective study", *Gynecologic Oncology*, Vol. 92, pp. 603–610, 2004.

61. Matorras, R., F. Matorras, R. Mendoza, M. Rodriguez, J. Remohi, F. J. Rodriguez-Escudero, and C. Simon, "The implantation of every embryo facilitates the chances of the remaining embryos to implant in an IVF programme: a mathematical model to predict pregnancy and multiple pregnancy rates.", *Human Reproduction*, Vol. 20, pp. 2932–2931, 2005.

62. Speirs, A., H. Baker, and N. Abdullah, "Analysis of factors affecting embryo implantation.", *Human Reproduction*, Vol. 11, pp. 187–191, 1996.

63. Trimarchi, J., "A mathematical model for predicting which embryos to transfer - an illusion of control or a powerful tool?", *Fertility and Sterility*, Vol. 76, pp. 1286–1287, 2001.

64. Racowsky, C., L. Ohno-Macado, J. Kim, and J. D. Biggers, "Is there an advantage in scoring early embryos on more than one day?", *Human Reproduction*, Vol. 29, No. 9, pp. 2104–2113, 2009.

65. Alpaydin, E., *Introduction to Machine Learning*, MIT Press, 2004.

66. Mladenic, D. and M. Grobelnik, "Feature selection on hierarchy of web documents", *Decision Support Systems*, Vol. 35, pp. 45–87, 2003.

67. Shlen, J., "A Tutorial on Principal Component Analysis", 2005, `http://www.cs.cmu.edu/~elaw/papers/pca.pdf`.

68. Lessmann, S., B. Baesens, C. Mues, and S. Pietsch, "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings", *IEEE Transactions on Software Engineering*, Vol. 34, pp. 485–496, 2008.

69. Viaene, S., R. Derrid, B. Baesens, and G. Dedene, "A Comparison of State-of-the-Art Classification for Expert Automobile Insurance Claim Fraud Detection", *The Journal of Rsik and Insurance*, Vol. 69, pp. 373–421, 2002.

70. Burges, C., "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121–167, 1998.

71. Rabiner, L., "A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition", *Prooceedings of the IEEE*, Vol. 77, pp. 257–286, 1989.

72. Guerif, F., 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", *Human Reproduction*, Vol. 22, pp. 1973–1981, 2007.

73. Heckerman, D., "A Tutorial on Learning With Bayesian Networks", Technical report, Microsoft Research Advanced Technology Division Microsoft Corporation, 1996.

74. Friedman, N., D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers", *Machine Learning*, Vol. 29, pp. 131–163, 1997.

75. Meloni, A., A. Ripoli, V. Positano, and L. Landini, "Mutual Information Preconditioning Improves Structure Learning of Bayesian Networks From Medical Databases", *IEEE Trans. On Information Technology In Biomedicine*, Vol. 13, pp. 984–989, 2009.

76. Lucas, P., "Restricted Bayesian Network Structure Learning", *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, pp. 217–232, Springer-Verlag, 2002.

77. Su, J., H. Zhang, C. Ling, and S. Matwin, "Discriminative Parameter Learning for Bayesian Networks", *25 th International Conference on Machine Learning (ICML)*, 2008.

78. Greiner, R. and W. Zhou, "Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers", *AAAI/IAAI*, 2002.

79. Damasevicius, R., "Optimization of SVM Parameters for Promoter Recognition in DNA Sequences", *International Conference, 20th EURO Mini Conference,Continuous Optimization and Knowledge-Based Technologies (EurOPT-2008)*, 2008.

80. Jung, T. and D. Polani, "Sequential Learning with LS-SVM for Large-Scale Data Sets", *ICANN*, 2006.

81. Bishop, C., *Pattern Recognition and Machine Learning*, Springer, 2006.

82. Johansson, S., M. Jern, and J. Johansson, "Interactive Quantification of Categorical Variables in Mixed Data Sets", *Proceedings of IEEE International Conference on Information Visualisation, IV08*, pp. 3–10, 2008.

83. Nukoolkit, C., H. Chen, and D. Brown, "A Data Transformation Technique for Car Injury Prediction", *Proceedings of the 39th Annual ACM-SE Conference*, 2001.

84. Moturu, S., H. Liu, and W. Johnson, "Healthcare Risk Modeling for Medicaid Patients", *International Conference on Healthcare Informatics*, pp. 126–133, Madeira, Portugal, January 2008.

85. Nicopoullos, J. D. M., C. Gilling-Smith, P. A. Almeida, S. Homa, L. Nice, H. Tempest, and J. Ramsay, "The role of sperm aneuploidy as a predictor of the success of intracytoplasmic sperm injection?", *Human Reproduction*, Vol. 23, pp. 240–250, 2007.

86. Marble, R. P. and J. C. Healy, "A neural network approach to the diagnosis of morbidity outcomes in trauma care", *Artificial Intelligence in Medicine*, Vol. 15, pp. 299–307, 1999.

87. Fawcett, T., "An Introduction to ROC Analysis", *Pattern Recognition Letters*, Vol. 27, pp. 861–874, 2006.

88. Huang, J. and C. Liang, "Using AUC and Accuracy in Evaluating Learning Algorithms", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, pp. 299–310, 2005.

89. Witten, I. H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2 edition, 2005.

90. Cohen, G., M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection", *Artificial Intelligence in Medicine*, Vol. 37, pp. 7–18, 2006.

91. Mazurowski, H. P., M.A. and, J. Zurada, J. Lob, J. Baker, and G. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", *Neural Networks*, Vol. 21, pp. 427–436, 2008.

92. Mac Namee, B., P. Cunningham, S. Byrne, and O. Corrigan, "The Problem of Bias in Training Data in Regression Problems in Medical Decision Support", *Artificial Intelligence in Medicine*, Vol. 24, pp. 51–70, 2002.

93. Uyar, A., A. Bener, H. N. Ciray, and M. Bahceci, "ROC Based Evaluation and Comparison of Classifiers for IVF Implantation Prediction", *Second International ICST Conference on Electronic Healthcare for the 21st century (eHealth)*, 2009.

94. Ciray, H. N., F. Bener, L. Karagenc, U. Ulug, and M. Bahceci, "Impact of Assisted Hatching on ART Outcome in Women with Endometriosis", *Human Reproduction*, 2005.

95. Netica., "Application for Belief Networks and Influence Diagrams", Norsys Software Corp., 1997.

96. Kohavi, R., P. Langley, and Y. Yun, "The Utility of Feature Weighting in Nearest-Neighbor Algorithms", *Proceedings of the Ninth European Conference on Machine Learning*, pp. 85–92, Springer-Verlag, 1997.

97. Vivencio, D., E. Hruschka, M. Nicoletti, E. dos Santos, and S. Galvio, "Feature-weighted k-Nearest Neighbor Classifier", *2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007)*, 2007.

98. Frank, A. and A. Asuncion, "UCI Machine Learning Repository", 2010, `http://archive.ics.uci.edu/ml`.

99. Janecek, A., W. Gansterer, M. Demel, and G. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy", *JMLR: Workshop and Conference Proceedings*, 2008.

100. Seli, E., D. Sakkas, R. Scott, S. Kwok, S. Rosendahl, and D. Burns, "Noninvasive metabolomic profiling of embryo culture media using Raman and near-infrared spectroscopy correlates with reproductive potential of embryos in women undergoing in vitro fertilization", *Fertility And Sterility*, Vol. 88, pp. 1350–1357, 2007.

101. Vergouw, C. G., L. L. Botros, P. Roos, J. W. Lens, R. Schats, P. G. A. Hompes, D. H. Burns, and L. C. B., "Metabolomic Profiling by Near-Infrared Spectroscopy

as a Tool to Assess Embryo Viability: A Novel, Non-Invasive Method for Embryo Selection", *Human Reproduction*, Vol. 23, pp. 1499–1504, 2008.

102. Pagidas, K., Y. Ying, and D. Keefe, "Predictive value of pre-implantation genetic diagnosis for aneuploidy screening in repeated IVF-ET cycles among women with recurrent implantation failure", *Journal of Assisted Reproduction and Genetics*, Vol. 25, pp. 103–106, 2008.

103. Wong, C., K. Loewke, N. Bossert, B. Behr, C. De Jonge, T. Baer, and R. Pera, "Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage", *Nature Biotechnology*, Vol. 28, No. 10, pp. 1115–1121, 2010.