Handling the Imbalance Problem of IVF Implantation Prediction

Asli Uyar *, Ayse Bener, H. Nadir Ciray, Mustafa Bahceci

Abstract— **Predicting implantation outcomes of in-vitro fertilization (IVF) embryos is critical for the success of the treatment. We have applied the Naive Bayes classifier to an original IVF dataset in order to discriminate embryos according to the implantation potentials. The dataset we analyzed represents an imbalanced distribution of positive and negative instances. In order to deal with the problem of imbalance, we examined the effects of oversampling the minority class, undersampling the majority class and the adjustment of the decision threshold on the classification performance. We have used features of Receiver Operating Characteristics (ROC) curves in the evaluation of experiments. Our results revealed that it is possible to obtain optimum True Positive and False Positive Rates simply by adjusting the decision threshold. Under-sampling experiments show that we can achieve the same prediction performance with less data as well as 736 embryo samples.**

Keywords: Implantation prediction, in-vitro fertilization, imbalance problem, Naive Bayes.

1 Introduction

Many real-world machine learning applications represent an imbalanced distribution of positive and negative classes where the number of instances in one class dominates the number of instances in the other class(es). In such cases, it is necessary to overcome any possible bias towards the majority class in the learning and prediction tasks. Consequently, learning from imbalanced datasets has been an important research interest in the last decade [1] [2]. Various sampling strategies have been proposed as a pre-processing stage in order to overcome the problem of imbalance [3] [4] [5]. On the other hand, it is also possible to adjust the decision threshold of classifiers in imbalanced datasets. Recent studies show that both approaches produce similar results [6] [7].

In this study, we focus on a specific area of medical diagnosis, i.e. in-vitro fertilization (IVF), to estimate the implantation potentials of embryos. When constructing predictive models in the IVF domain, the input data consist of a set of prognostic factors obtained from retrospective clinical databases and generally contain fewer samples with positive outcomes. Any classifier built on these datasets has much more information to identify unsuccessful IVF treatments in comparison to successful ones. Therefore, the implantation prediction is handled as a typical case of learning from imbalanced data problems. We analyzed the effects of resampling the training data and decision threshold optimization on imbalanced IVF dataset using the Naive Bayes classifier. Our results show that 0.3 is the best threshold for classifying the implantation outcomes of embryos.

We have also considered another research problem, which consists of the determination of the smallest amount of training data required to build an effective predictor model. Data collection is a costly and time-consuming process in medical applications. The analysis of undersampling experiments led us to define the sufficient size of embryo samples for implantation prediction that would reduce the effort spent for data collection in the IVF domain. This paper is an extended version of the work presented in [8].

The rest of the paper is organized as follows: Section 2 describes the IVF domain with an emphasis on the implantation prediction and the characteristics of the IVF dataset. Brief definitions of the Naive Bayes classifier, ROC curves and sampling strategies are given in Section 3. Section 4 represents the experiments and the results. Finally, we conclude in Section 5 with a discussion on the results.

2 In-Vitro Fertilization

Infertility is defined as a couple's biological inability to conceive 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 a baby

*A. Uyar is with the Department of Computer Engineering, Bogazici University, 34342 Bebek Istanbul Turkey (phone: +90-212-359-7227; fax:+90-212-287-2461; e-mail: asli.uyar@boun.edu.tr)

A. Bener is with the Department of Computer Engineering, Bogazici University, 34342 Bebek Istanbul Turkey (e-mail: bener@boun.edu.tr)

H. N. Ciray is with Bahceci Women Health Care Centre, Azer Is Merkezi 44/17 Kat 6, Abdi Ipekci Cad., 80200 Nisantasi, Istanbul, Turkey (e-mail: nadirc@superonline.com)

M. Bahceci is with Bahceci Women Health Care Centre, Azer Is Merkezi 44/17 Kat 6, Abdi Ipekci Cad., 80200 Nisantasi, Istanbul, Turkey (e-mail: mbahceci@superonline.com)

This study is supported by Bahceci Women Health Care Centre and by Bogazici University, (BAP) under grant number 09A104D.

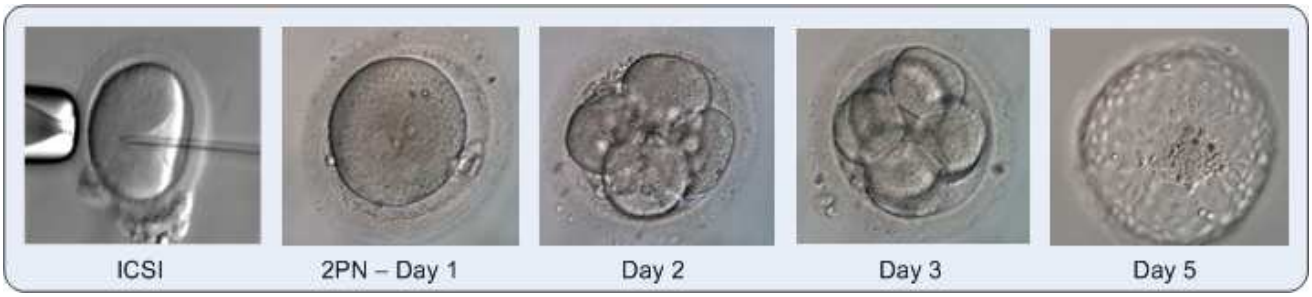

Figure 1: Human germ cells, Intra-Cytoplasmic Sperm Injection (ICSI)(ICSI is a method during which a single sperm cell is injected into the cytoplasm of the oocyte) and embryo growth day by day

spontaneously. Once the infertility factor of a couple is determined, an appropriate assisted reproduction treatment is applied in order to conceive a successful pregnancy.

IVF [1] is a common infertility treatment method during which female germ cells (oocytes) are inseminated by sperm under laboratory conditions. The fertilized oocytes are cultured between 2-6 days in special medical equipments, and embryonic growth is observed and recorded by embryologists. Finally, the selected embryo(s) are transferred into the woman's womb. Figure 1 represents images emphasizing the IVF procedure and the embryo morphology observed day by day in IVF laboratories.

The selection of the embryos with the highest reproductive potential and the decision concerning the number of embryos to be transferred are crucial for achieving successful pregnancy. Predicting implantation potentials of individual embryos may expedite and enhance expert judgement for these critical decisions. Implantation is defined as the attachment of the embryo to the inner layer of the womb, and a positive implantation outcome is defined as the ultrasound visualization of a pregnancy sac with fetal heart activity at 12 weeks after embryo transfer.

2.1 Implantation Prediction

This study focuses on predicting the implantation outcomes of IVF embryos. Obtaining multiple embryos at each cycle of the treatment is possible. Embryologists decide which embryos to transfer and which ones to freeze or further culture (Figure 2). These decisions are mostly based on clinical traditions and the personal experience of embryologists. However, making future predictions about embryo growth is a challenging process depending on various embryo and patient related variables.

The accurate prediction of implantation potentials is especially crucial for determining the number of embryos to be transferred. Generally, the three highest quality embryos at the most are transferred to a woman's uterus. Multiple embryo transfers increase the pregnancy prob-

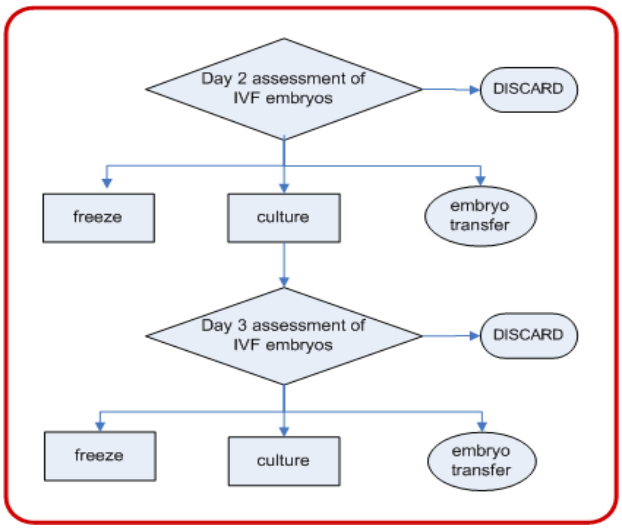

Figure 2: Critical decisions after day 2 and 3 evaluation of IVF embryos (Modified from [13])

ability, but they also increase possible complications of multiple pregnancies for both the mother and fetuses [9] [10]. In some cases, a multifetal pregnancy reduction may be necessary in order to decrease the number of fetuses in multiple pregnancies. This operation is expected to reduce the risk of miscarriage and premature birth and to increase the chance of survival for the remaining fetus(es). However, reduction may have a negative psychological impact on the parents. Therefore, the prevention of IVF multiple pregnancies should be preferred over multifetal pregnancy reduction.

Elective single embryo transfer (eSET) has been favored as a solution to the IVF multiple pregnancy problem [11]. However, a recent survey on perceived barriers to eSET among IVF professionals reported that 47% of IVF professionals refuse the use of eSET due to uncertainty about the technique or lack of prognostic factors and models for determining eSET candidates [12].

For applicability in clinical practice, physicians need reliable eSET criteria depending on two main issues: the selection of the most viable embryos and the identifica-

tion of patients suitable for eSET. Therefore, objective predictor models are required to predict the implantation potentials of embryos related to the characteristics of both the embryo and the patient. From the perspective of machine learning, implantation prediction is considered as a binary (2-class) classification problem where the classes represent positive and negative implantation outcomes.

2.2 Machine Learning in the IVF Domain

IVF treatment is a complex and costly process requiring continuous observation and critical decisions of embryologists in certain stages. Contrary to the importance and emergence of intelligent decision support systems in the IVF process, the related literature is limited. Existing studies mostly deal with predicting the pregnancy outcome of the IVF cycle (that is either multiple pregnancy or single pregnancy) rather than predicting the implantation outcome of individual embryos.

As shown by preliminary studies, a case-based reasoning system [14] and neural networks have been constructed in predicting the outcome of in-vitro fertilization [15]. Later, decision tree models were applied for predicting the pregnancy outcome from clinical IVF data [16][17]. The most recent study on implantation prediction proposes a Bayesian classification system for embryo selection [18]. The direct comparison of these studies and the presented results are not possible due to the diverseness of research objectives, input feature sets of data, training and testing strategies, and performance measures.

Most studies presenting predictive models in the IVF domain suffer from insufficient results [15][19][20][18]. One of the reasons for poor prediction performance may be the limited number of data samples. Thus, performing experiments on larger datasets may be necessary. However, the acquisition of complete and reliable medical data is a challenge for machine learning researchers. Therefore, it is crucial to determine the minimum number of training samples in order to prevent much effort from being wasted during the collection of data.

2.3 Dataset

Certain legislative rules have been defined related to IVF treatment in every country due to social ethical reasons. Usually, the restrictions apply to donation, embryo manipulation, the number of embryos to be transferred in each cycle etc. In addition to legal procedures, each IVF clinic implements different technologies and methodologies in practice. Because of this variety, IVF clinics have distinctive databases, and unfortunately there are no public IVF datasets in the machine learning community. In this study, we analyzed the IVF procedure and the related database of Bahceci Women Healthcare Centre in Istanbul.

Table 1: Selected dataset features for each embryo feature vector

Dataset Features	**Data Type**
Patient Characteristics	
Age of female patient	Numerical
Infertility factor	Categorical
Treatment protocol	Categorical
Follicular stimulating hormone dosage	Numerical
Peak Estradiol level	Numerical
Embryo Morphological Data	
Early cleavage morphology	Categorical
Early cleavage time	Numerical
Number of cells	Numerical
Nucleus characteristics	Numerical
Fragmentation rate	Numerical
Equality of blastomeres	Numerical
Appearance of cytoplasm	Categorical
Transfer Data	
Transfer day	Categorical
Physician performing embryo transfer	Categorical
Difficulty of transfer	Categorical

Initially, a dataset from an existing IVF database, which included individual embryo feature vectors, was constructed. Each embryo was represented with 15 variables (Table 1), and a class label was assigned: +1 and -1 indicating that the implantation was successful or not successful, respectively. A positive implantation outcome was defined as fetal cardiac activity at 12 weeks following embryo transfer. Dataset features and data types are given in Table I. The features have been selected depending on the experiences of senior embryologists in the clinic and the related studies in the literature [18]. Apart from the existing studies, we have also considered the influence of the physician performing the embryo transfer [21] and the difficulty of the transfer [22] as prognostic factors.

Input data features include both continuous (e.g. age, hormone levels etc.) and categorical (infertility factor, treatment protocol etc.) variables. The IVF 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 and 331 (14.6%) positive implantation outcomes. Hence, implantation prediction is handled as a typical case of learning from imbalanced data problem.

3 Methodology

In a previous study, we have compared various classifiers for the implantation prediction of IVF embryos and shown that the Naive Bayes classifier produces a significantly better predictive performance [23]. Therefore, we apply the Naive Bayes algorithm to imbalanced IVF dataset in order to investigate the effect of sampling strategies and threshold optimization.

This section briefly describes the Naive Bayes classifier, performance measures related to ROC analysis and the problem of learning from imbalanced datasets.

3.1 Naive Bayes Classification

The Bayes theorem given below states that the posterior probability of a sample $P(C_i|x)$ is related to the prior distribution $P(x|C_i)$ and the likelihood $P(C_i)$ [24].

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)} \quad (1)$$

According to Bayes decision theory, a sample x is said to belong to class C_j with the highest posterior probability $C_j = \max_i(P(C_i|x))$.

3.2 ROC Analysis and Performance Criteria

In the machine learning community, after the detection of the weakness of simple error rate as a performance measure, the use of ROC curves has gained increasing attention [25]. In this study, we use ROC curves to evaluate the discriminative performance of the binary Naive Bayes classifier where each instance I is mapped to one of the positive and negative classes labeled as *+1* and *-1* respectively. Given a classifier and an instance, the prediction outcomes depending on actual class labels of instances can be represented as a 2x2 confusion matrix as shown in Table 2.

Table 2: Confusion Matrix

	Predicted	
Actual Case	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Common classifier performance metrics have been derived from the confusion matrix:

- ***TP rate (TPR)*** is a measure of accuracy for correctly detecting the positive instances and is equal to the ratio of number of true positives (TP) over the sum of true positives and false negatives (FN). TPR (also called as Hit Rate) corresponds to *sensitivity* in medical diagnosis.

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

- ***FP Rate (FPR)*** represents the number of false alarms that is the false positives (FP) over the sum of true negative (TN) and false positives (FP). FPR corresponds to (*1 - specificity*) in medical domain.

$$FPR = \frac{FP}{TN + FP} \quad (3)$$

It is necessary to mention the critical points on the 2D ROC curve. The lower left point (0,0) represents the assigning of all instances to negative class. Hence, there are no positive predictions yielding TPR and FPR to be 0. Conversely, the 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}).

3.3 The Problem of Imbalanced Datasets

In classification tasks, when the aim of the classification is to maximize accuracy, imbalanced datasets produce an unsatisfactory prediction performance. For example, in the IVF dataset we used, any classifier labeling each instance with the negative class will achieve 84.6% accuracy. However, it will actually produce 0% TPR. In such cases, the desired solution is to find an acceptable trade-off between the TPR and the FPR of the classification.

3.3.1 Sampling

A common approach for overcoming the imbalance problem is to rebalance the datasets artificially. Two main sampling strategies are oversampling, which replicates instances from the minority class [4] and the undersampling, where some of the instances in the majority class are removed [3].

The effects of sampling methods in prediction performance have been investigated in machine learning based medical decision-making applications [26] [27] [28]. We have performed oversampling and undersampling in different scales and examined the classification performance on the rebalanced IVF data with the default threshold of 0.5.

3.3.2 Threshold Optimization

It is also necessary to investigate the effect of adjusting the output threshold for a particular classifier. Many machine learning algorithms (i.e. Naive Bayes) produce an estimate of the probability of class membership for a binary classification problem. When using the Naive Bayes classifier, the TPR and FPR have been calculated for a single decision threshold (default: 0.5) that maps to a single point on the ROC curve.

However, Provost clearly defined that applying the standard machine learning algorithms to imbalanced datasets without adjusting the decision threshold may be a critical mistake [7]. Therefore, it is necessary to evaluate the performance of classification for different thresholds since it would be sufficient to find the optimum threshold rather than changing the balance ratio of dataset.

Table 3: Distribution of classes and prediction results after oversampling the training data.

Dataset No	1	2	3	4	5	6	7	8	9	10
Number of Positive Instances	218	436	654	872	1090	1308	1526	1744	1962	2180
Number of Negative Instances	1295	1295	1295	1295	1295	1295	1295	1295	1295	1295
True Positive Rate	0.508	0.630	0.665	0.692	0.705	0.723	0.741	0.749	0.760	0.768
False Positive Rate	0.180	0.287	0.336	0.372	0.404	0.429	0.449	0.461	0.473	0.488

Table 4: Distribution of classes and prediction results after undersampling the training data.

Dataset No	1	2	3	4	5	6	7	8	9	10
Number of Positive Instances	218	218	218	218	218	218	218	218	218	218
Number of Negative Instances	1295	1165	1036	906	777	647	518	388	259	129
True Positive Rate	0.508	0.542	0.554	0.581	0.611	0.637	0.653	0.682	0.726	0.791
False Positive Rate	0.180	0.202	0.22	0.245	0.262	0.298	0.321	0.360	0.414	0.513

4 Experiments and Results

We have conducted experiments to investigate the effects of oversampling and undersampling the IVF data and moving the decision threshold of the Naive Bayes classifier for the implantation prediction problem. Classification experiments were performed using the Weka data mining tool [29].

4.1 Training and Testing Strategy

Two-thirds of the dataset were randomly selected for establishing a predictor model and the remaining one-third was utilized for testing. This initial random splitting was performed using the stratification principle in order to ensure that the proportion of positive and negative classes remained the same in both the training and the testing sets as in the original dataset. Then, the distribution of the training data has been artificially changed.

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

For both sampling methods, the trained model was tested on the separate 1/3 dataset including a total of 762 embryo records with 649 negative and 113 positive implantation outcomes. The random two-thirds to one-third partitioning of the dataset into training and testing sets was repeated 10 times in order to overcome sampling bias. Oversampling and undersampling processes were repeated for each of the 10 holdout experiments. The presented results are the mean of these 10 repetitions.

4.2 Results

Table 3 and Table 4 represent the distribution of the training set and prediction results in terms of TPR and FPR for oversampling and undersampling 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 or reducing the number of negative embryo samples raises the number of positive predictions. The trade-off between the TPR and the FPR can be adjusted by changing the ratio of classes.

An optimum (TPR, FPR) pair can be obtained as explained in Section 3.2. These correspond to (66.5%, 33.6%) and (65.3%, 32.1%) for oversampling and undersampling, respectively. Undersampling experiments 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.

The TPR and FPR values were been calculated by varying the decision thresholds in the range of [0:0.1:1]. The resulting set of (TPR, FPR) pairs are given in Table 5.

The results of oversampling, undersampling and threshold variation have been plotted as a single 2D ROC curve (Figure 3). Both sampling methods and the adjustment of the decision threshold produce almost the same ROC curve, demonstrating the similarity of the effects of these methods on prediction performance.

Classification with the default decision threshold, i.e. 0.5, produces 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 curve reduces the FPR, but often results in a lower TPR as well. The thresholds on the right hand side increase both the TPR and the FPR.

4.3 Threats to Validity

In machine learning applications, it is crucial to deal with possible biases arising from sampling procedure and the training/testing strategies. In order to overcome sampling bias, we have applied ten repetitions of the random

Table 5: Prediction results depending on the 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	0.771	0.694	**0.644**	0.584	0.508	0.413	0.280	0.131	0.036	0
False Positive Rate	1	0.482	0.376	**0.306**	0.238	0.180	0.131	0.086	0.046	0.006	0

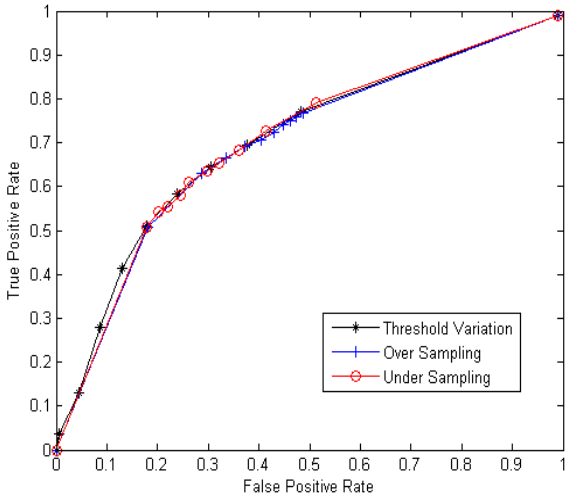

Figure 3: ROC curves demonstrating the effect of sampling and threshold variation of Naive Bayes based IVF implantation prediction

train/test set partitioning. In terms of construct validity, our observations are well translated into measures such as TPR and FPR measures that are clear and widely accepted by researchers for imbalanced datasets. The data come from a single source challenging the external validity of the results. However, public datasets do not exist in this domain, and other labs are reluctant to share their data.

5 Conclusions

Each real-world application of standard machine learning algorithms requires careful analysis of the input data and the utilized methods. Selecting the most appropriate pre-processing or post-processing tasks provides better prediction performance. This is crucial for providing reliable decisive support to domain experts especially in medical decision-making applications.

Most medical datasets represent an imbalanced distribution of positive and negative samples. This study has investigated the problem of learning from imbalanced datasets for the specific IVF domain. We have examined the effects of sampling and threshold optimization in Naive Bayes classification and presented a comparative analysis of these methods for the implantation prediction of IVF embryos.

Experimental results 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 conclude that it is not necessary to artificially rebalance the distribution of class samples in the IVF dataset.

An easier and more effective way is to find the optimum decision threshold that produces the required TPR and FPR values depending on the cost of misclassifications. Assuming that the the costs of false positive and false negative errors are equal, the optimum decision threshold is found to be 0.3 resulting in 64.4% TPR and 30.6% FPR in the implantation prediction. Furthermore, the analysis of the classification results related to the rebalanced datasets provided the minimum number of data instances required to train a predictor model in the implantation prediction problem.

References

[1] K. Huang, 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.

[2] L. Mena and J. Gonzalez, "Machine learning for imbalanced datasets: Application in medical diagnostic," in *19th International FLAIRS Conference (FLAIRS-2006)*, Melbourne Beach, Florida, May 11-13 2006.

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

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

[5] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Syntethic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[6] A. M. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," *Workshop on Learning from Imbalanced Data Sets*, 2003.

[7] F. Provost, "Machine learning from imbalanced data sets 101," in *Working Notes AAAI00 Workshop Learning from Imbalanced Data Sets*, 2000, pp. 1–3.

[8] A. Uyar, A. Bener, H. Ciray, and M. Bahceci, "Predicting implantation outcome from imbalanced IVF dataset," *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2009, WCECS 2009, 20-22 October 2009, San Francisco, USA*, pp. 803–808.

[9] J. Gerris and D. De Neubourg, "Single embryo transfer after IVF/ICSI: present possibilities and limits," *J Obstet Gynecol India*, vol. 55, pp. 26–47, 2005.

[10] A. Thurin, 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," *N Engl J Med*, vol. 351, pp. 2392–402, 2004.

[11] H. Martikainen, 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.

[12] A. van Peperstraten, 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.

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

[14] I. Jurisica, 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] S. J. Kaufmann, 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.

[16] R. Saith, A. Srinivasan, D. Michie, and I. 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] J. R. Trimarchi, J. Goodside, L. Passmore, T. Silberstein, L. Hamel, and L. Gonzalez, "Comparing data mining and logistic regression for predicting IVF outcome," *Fertil. Steril.*, 2003.

[18] D. A. Morales, 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.

[19] G. Venkat, R. Al-Nasser, S. Jerkovic, and I. Craft, "Prediction of success in IVF treatments using neural networks," *Fertility and Sterility*, vol. 82, p. 215, 2004.

[20] L. D. M. Ottosen, U. Ulrik Kesmodel, J. Hindkjr, and H. J. Ingerslev, "Pregnancy prediction models and eSET criteria for IVF patients do we need more information?" *J Assist Reprod Genet*, vol. 24, pp. 29–36, 2007.

[21] A. Angelini, "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.

[22] C. Tomas, 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.

[23] A. Uyar, A. Bener, H. Ciray, and M. Bahceci, "ROC based evaluation and comparison of classifiers for IVF implantation prediction," in *In Proc. Second International ICST Conference on Electronic Healthcare for the 21st century*, 2009.

[24] E. Alpaydin, *Introduction to Machine Learning.* MIT Press, 2004.

[25] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[26] G. Cohen, 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.

[27] H. P. Mazurowski, 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.

[28] B. Mac Namee, 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.

[29] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.