Predicting Implantation Outcome from Imbalanced IVF Dataset

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 Naive Bayes classifier to an original IVF dataset in order to discriminate embryos according to 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 over sampling the minority class, under sampling the majority class and 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 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 that of the other. In such cases, it is necessary to overcome 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 to deal with the problem of imbalance [3] [4] [5]. 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 [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 IVF domain, the input data consists 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 compared to successful ones. Therefore, implantation prediction is handled as a typical case of learning from imbalanced data problem. We analyze the effects of re-sampling the training data and decision threshold optimization on imbalanced IVF dataset using Naive Bayes classifier. Our results show that 0.3 is the best threshold for classification of embryos.

We have also considered another research problem that is 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. Analysis of under-sampling experiments leaded to define sufficient size of embryo samples for implantation prediction that would reduce the effort spent for data collection in IVF domain.

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

2 In-Vitro Fertilization

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. 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. Fertilized oocytes are cultured between 2-6 days in special medical equipments and embryonic growth is observed and recorded by embryologists. (Figure 1 represents images to give emphasize on IVF procedure and embryo morphology.) Finally, selected embryo(s) are transferred into the woman's womb. Selection of the embryos with highest reproduc-

*Department of Computer Engineering, Bogazici University, 34342 Bebek Istanbul Turkey, Email: asli.uyar@boun.edu.tr This study is supported by Bahceci IVF Centre and by Bogazici University, (BAP) under grant number 09A104D.

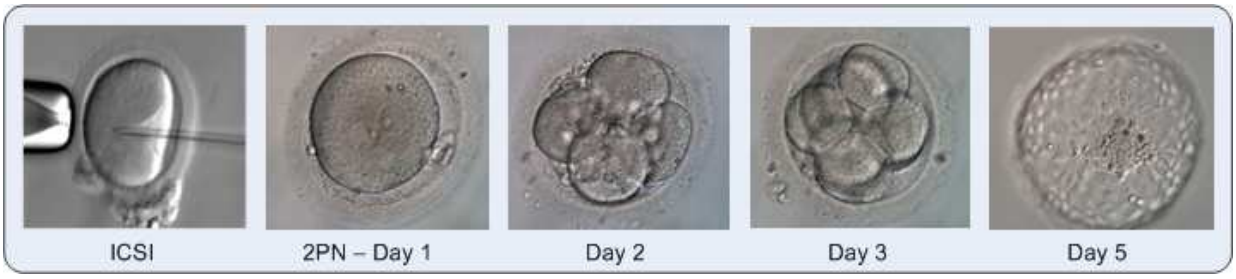

Figure 1: Human germ cells, Intra-Cytoplasmic Sperm Injection (ICSI)(ICSI is a method during which a single sperm cell is injected into the cytoplasm of the oocyte) and embryo growth day by day

tive potentials and the decision of number of embryos to be transferred is crucial for achieving successful pregnancies. Predicting implantation (i.e. attachment of the embryo to the inner layer of the womb) potentials of individual embryos may expedite and enhance expert judgement for these critical decisions.

2.1 Implantation Prediction

This study is concentrated on predicting implantation outcomes of IVF embryos. At each cycle of 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 [8] [9]. Elective single embryo transfer (eSET) has been favored as a solution to IVF multiple pregnancy problem. To be applicable in clinical practice, physicians need reliable eSET criteria depending on two main issues: selection of the most viable embryos and identification of patients suitable for eSET. Therefore, objective predictor models are required to predict implantation potentials of embryos related to both embryo and patient characteristics. From the machine learning point of view implantation prediction is considered as a binary (2-class) classification problem where the classes represent positive and negative implantation outcomes.

2.2 Related Work

IVF treatment is a complex and costly process requiring continuous observation and critical decisions of embryologists in certain stages. On the contrary to the importance and emergence of intelligent decision support systems in IVF process, the related literature is limited. As the preliminary studies, a case-based reasoning system [10] and neural networks have been constructed in predicting the outcome of in-vitro fertilization [11]. Later, decision tree models were applied for prediction of pregnancy outcome from clinical IVF data [12][13]. The most recent study on implantation prediction proposes a Bayesian classification system for embryo selection [14]. Direct comparison of the presented results is not possible due to variety of research objectives, input feature sets of data, training and testing strategies and performance measures.

Table 1: Selected dataset features for each embryo feature vector

Dataset Features	**Data Type**
Patient Characteristics	
Woman age	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

Most studies presenting predictive models in IVF domain suffer from insufficient results [11][15][16][14]. One of the reasons for poor prediction performance may be limited number of data samples and it may be necessary to perform experiments on larger datasets. However, acquisition of complete and reliable medical data is a challenge for machine learning researchers. Therefore, it is crucial to determine minimum number of training samples in order to prevent waste of effort spent on data collection.

2.3 Dataset

Because of social and ethical reasons in every country some legislative rules have been defined related to IVF treatment. Usually, the restrictions apply for donation, embryo manipulation, number of embryos to be transferred in each cycle etc. Besides the legal procedures, each IVF clinic applies 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 analyze the IVF procedure and related database of IVF Unit of German Hospital in Istanbul.

Initially, a dataset from an existing IVF database was constructed which included individual embryo feature vectors. Each embryo was represented with 15 variables (Table 1) and a class label was assigned: +1 and -1 indicating that implantation was successful or not-successful, respectively. A positive implantation outcome was defined as foetal cardiac activity at 12 weeks after embryo transfer. Dataset features and data types are given in Table I. The features have been selected depending on experiences of senior embryologists in the clinic [17] and related studies in the literature [14]. Apart from existing studies, we have also considered the effect of physician performing embryo transfer [18] and the difficulty of transfer [19] 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 implantation 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 implantation prediction of IVF embryos and shown that Naive Bayes produce significantly better predictive performance [20]. Therefore, we apply 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

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

$$P(C_i|x) = P(x|C_i)P(C_i)/P(x) \quad (1)$$

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 realization of the weakness of simple error rate as a performance measure, the use of ROC curves have gained an increasing attention [21]. In this study, we use ROC curves to evaluate the discriminative performance of 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 Hit Rate) corresponds to *sensitivity* in medical diagnosis.

$$TPR = (TP)/(TP + FN) \quad (2)$$

- ***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.

$$FPR = (FP)/(TN + FP) \quad (3)$$

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}).

3.3 The Problem of Imbalanced Dataset

In classification tasks, when the aim of the classification is to maximize the accuracy, imbalanced datasets produce unsatisfactory prediction performance. For example, in the IVF dataset we used, any classifier labeling every 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 tradeoff between TPR and FPR of classification.

3.3.1 Sampling

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 [4] and under-sampling where some of the instances in the majority class is removed [3]. The effects of sampling methods in prediction performance have been investigated in machine learning based medical decision making applications [22] [23] [24]. We have performed over-sampling and under-sampling in different scales and examined the classification performance on the re-balanced IVF data with the default threshold of 0.5.

3.3.2 Threshold Optimization

It is also necessary to investigate the effect of adjustment of 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 Naive Bayes classifier, 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, it may be a critical mistake to apply the standard machine learning algorithms to imbalanced datasets without adjusting the decision threshold [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.

4 Experiments and Results

We have conducted experiments to investigate the effects of over-sampling and under-sampling the IVF data and moving the decisison threshold of Naive Bayes classifier for implantation prediction problem. Classification experiments have been performed using Weka data mining tool [25].

4.1 Training and Testing Strategy

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

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.

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, one-third partitioning of dataset into training and test sets has been repeated 10 times in order to overcome sampling bias. Over sampling and under sampling processes have been repeated for each of the 10 hold out 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 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 be obtained as explained in Section 3.2. These corresponds to (66.5%, 33.6%) and (65.3%, 32.1%) for over sampling and under sampling, respectively. Under sampling 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 have also 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 over-sampling, under-sampling and threshold variation have been plotted as a single 2D ROC curve (Figure 2). 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.

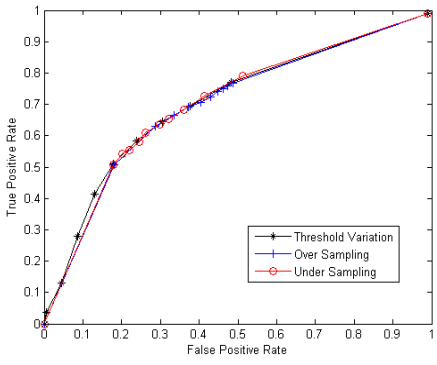

Figure 2: ROC curves demonstrating the effect of sampling and threshold variation of Naive Bayes based IVF implantation prediction

Table 3: Distribution of classes and prediction results after over sampling 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 under sampling 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

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

4.3 Threats to Validity

In machine learning applications, it is crucial to deal with possible biases arising from sampling procedure and training-testing strategies. In order to overcome sampling bias, we have applied ten repetitions of the random 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 comes from a single source challenging the external validity of the results. However, in this domain there are no public datasets nor different labs are willing to share their data.

5 Conclusions

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. This study has investigated the problem of learning from imbalanced dataset for the specific IVF domain. We examined the effects of sampling and threshold optimization in Naive Bayes classification and presented a comparative analysis of these methods for 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 rebalancing 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. Assuming equal cost of false positive and false negative errors, the optimum decision threshold is found to be 0.3 resulting in 64.4% TPR and 30.6% FPR in implantation prediction. Furthermore, analysis of the classification results on rebalanced datasets provided the minimum number of data instances required to train a predictor model in 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.

Table 5: Prediction results depending on variation of 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

[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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] H. N. Ciray, S. Tosun, O. Hacifazlioglu, A. Mesut, and M. Bahceci, "Prolonged duration of transfer does not affect outcome in cycles with good embryo quality," *Fertil. Steril.*, 2007.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

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