Gynecology & Obstetrics
ISSN: 2161-0932

John et al., Gynecol Obstet (Sunnyvale) 2019, 9:2
DOI: 10.4172/2161-0932.1000497

Case Report Open Access

An Application of Machine Learning in IVF: Comparing the Accuracy of Classification Alogithims for the Prediction of Twins

Rinehart John*

Reproductive Medicine Institute, Evanston, IL, United States

***Corresponding author**: Rinehart John, Reproductive Medicine Institute, Evanston, IL, United States, Tel: +847-869-7777; E-mail: jsrinehart@aol.com

Received date: January 13, 2019; **Accepted date**: February 01, 2019; **Published date**: February 07, 2019

Abstract

Background: Clinical decision-making dilemmas are particularly notable in IVF practice, given that large datasets are often generated which enable clinicians to make predictions that inform treatment choices. This study applied machine learning by using IVF data to determine the risk of twins when two or more embryos are available for transfer. While most classifiers are able to provide estimates of accuracy, this study went further by comparing classifiers both by accuracy and Area Under the Curve (AUC).

Methods: Study data were derived from a large electronic medical record system that is utilized by over 140 IVF clinics and contained 135,000 IVF cycles. The dataset was reduced from 88 variables to 40 and included only those cycles of IVF where two or more blastocyst embryos were created. The following classifiers were compared in terms of accuracy and AUC: a generalized linear model, linear discriminant analysis, quadratic discriminant analysis, K-nearest neighbors, support vector machine, random forests, and boosting. A stacking ensemble learning algorithm was also applied in order to use predictions from classifiers to create a new model.

Results: While the ensemble classifier was the most accurate, none of the classifiers predominated as being significantly superior to other classifiers. Findings indicated that boosting methods for classifiers performed poorly; logistic and linear discriminant analysis classifiers performed better than the quadratic discriminant analysis classifier, and the support vector machine performed almost as well as the tree classifier. AUC results were consistent with the comparisons for accuracy. External validation was also performed using a different dataset containing 588 observations. All models performed better using the external validation dataset, with the random forest classifier performing markedly better than any other classifier.

Conclusions: These results support the impression that big data can be of value in the clinical decision-making process; but that no single statistical algorithm provides maximum accuracy for all databases. Therefore, different datasets will require investigation in order to determine which algorithms are the most accurate for a particular set of data. These findings underscore the premise that clinicians with access to large amounts of data can use advanced predictive analytic models to create robust clinical information of vital importance for patient care.

Keywords: *In-vitro* fertilization (IVF); Machine learning; Classification algorithms; Artificial intelligence (AI); Assisted reproductive technology (ART); Ensemble learning; Linear classifiers; Tree-based classifiers; Stacking

Introduction

The digital revolution permitted unimaginable amounts of data to be retained and accessed. These vast amounts of data contain considerable information. The field of Artificial Intelligence (AI) emerged so that the information stored in the data could be obtained and used in a variety of settings. Machine learning is a subset of AI [1] that enables researchers to learn from data via automated model building and minimal human intervention. Advances in model construction have given the researcher many options for evaluating a set of data. Early models frequently used linear regression; but as new technologies were introduced, the tool kit expanded accordingly. Many of the clinical challenges faced by physicians are actually classification problems. While classifiers may be based upon regression methods; they may also be derived from next-nearest neighbor models, support vector machines, tree-based models, or neural networks-among many others [2]. The choice of which model to use becomes important since no one model fits all datasets and models based upon a training set may not be accurate when applied to a specific set of data [3]. Applying various classifiers to a set of data to answer a simple question exemplifies the dilemma of how to select the best model for the problem in question.

One approach for creating classification algorithms utilizes ensemble learning [4]. Ensemble learning is based upon the concept that, although a single algorithm may represent a weak predictor; combining multiple algorithms results in much more accurate predictions. While a number of ensemble techniques are currently available, the present study evaluated tree-based methods and stacking.

Tree-based classification is a type of ensemble learning that has the advantages of interpretability, ease-of-use, and accuracy [5]. Because decision trees are similar to flowcharts-which are frequently used for medical decision-making, they are readily understandable to the practicing clinician. Furthermore, trees can be displayed visually for ease of interpretation and are simply created using desktop programs and data from electronic medical records. When applied to a dataset,

trees may be more accurate than linear regression classification methods. As they require neither normalization nor standardization, decision trees are easier to use relative to linear models [6]. The ease of data input offered by decision trees makes them especially advantageous to clinicians. A final advantage of decision tree software is its ability to effectively handle missing data. Given that they are not concerned with outliers, tree-based models require minimal data preparation. Each of these factors represents a critical influence on whether a model is accepted for use in a clinical setting. Importantly, clinical acceptance-which is a major obstacle of predictive analytic utilization, is enhanced in models regarded as high in ease of use.

Stacking is another type of ensemble machine learning in which different classifiers are combined such that the output of one classifier may be used as the input of another classifier [4]. All classifiers use the same dataset such that each classifier is independent of other classifiers. Stacking uses the predictions from classifiers to create a new model. As submodels created in stacking are not required to be the best models, it is not necessary to fine-tune each model; but rather, to show an increase in model accuracy over the baseline prediction [7].

The purpose of this study was to use a clinical problem from IVF to compare the accuracy of numerous classification algorithms including ensemble learning methods. The clinical problem chosen was to predict the risks of twins when two or more embryos are available for transfer. The accuracy of the prediction was not the subject of this study. Most classifiers produce an estimate of accuracy; but, for many datasets, the Area Under the Curve (AUC) provides another method to determine which model is most accurate [8]. Therefore, the present study compared classifiers both by accuracy and AUC.

Methods

Data

The data was generously provided by PracticeHwy, which owns an electronic clinical software system (eIVF) designed specifically for IVF. Created by PracticeHwy for use by investigators, more than 140 IVF programs currently use the eIVF software [9]. The data was provided as an Excel spreadsheet consisting of input entered from various IVF programs. The initial database included 88 variables and 138,526 observations, where each observation represented one cycle of IVF. Following the initial evaluation of the variables based upon the expertise of the author and the issue addressed by this study, the number was narrowed to 40 variables. The final dataset used herein included only those cycles of IVF where two or more blastocyst embryos were created. Only when two or more blastocyst embryos exist will there be a need to decide whether one or two embryos should be transferred. Median imputation was used for missing data.

Statistical analysis

Eight different classifiers were used on a common training set. The models created by the classifiers were then applied to a common test set to determine which classifier, if any, was superior. Model performance was evaluated using accuracy and AUC. The following classifiers were included: Generalized Linear Model (GLM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), random forests, and boosting.

A second ensemble learning algorithm used stacking for model formation. Stacking is a process of combining different classifiers such that the output of one classifier may be used as the input of another classifier. All classifiers use the same dataset so that each classifier is independent of the other classifiers. Stacking uses the predictions from classifiers to create a new model. The program Super Learner [10] in r was used for the ensemble classifier using 6 classifiers and the common dataset that was used for all classifiers. The model creates a weighted average using the predictions from the classifiers. Super Learner uses nested cross-validation to assess the performance of the ensemble algorithm.

Results

As presented in Table 1, a comparison of the accuracy of classification algorithms identifies tree-based models as the most accurate classifier. However, none of the classifiers predominated as being significantly superior to other classifiers. The boosting method for classification performed poorly, which is not consistent with the overall impression that tree-based classifiers are better.

This finding demonstrates that there may be significant differences amongst the various classifiers.

Just as indicated from the comparison of the three tree-based classifiers, the comparison of the three linear regression-based classifiers demonstrates a difference in performance.

Both the logistic and linear discriminant analysis classifiers performed better than the quadratic discriminant analysis classifier.

The SVM performed almost as well as the tree classifier, which is consistent with the development of SVM analysis that was introduced in the 1990s as an improved leaning machine for classification problems [11].

Parameter	Linear classifiers					Tree-based classifiers		
Program	GLM	LDA	QDA	KNN	SVM	TREE	FOREST	BOOST
Accuracy	0.7534	0.7537	0.7131	0.7368	0.7545	0.7549	0.754	0.5939

Table 1: Comparison of the accuracy of eight classification algorithms.

The classifiers were also compared based upon the AUC using ROC-based analysis, as shown in Table 2.

The results from the comparison of the AUC mimic the comparisons for accuracy. Interestingly, these results provide an example of the accuracy and the AUC as discordant for the boosting classifier.

Program	GLM	LDA	QDA	KNN	SVM	TREE	FOREST	BOOST
	0.5855	0.5845	0.53	0.5534	0.5468	0.6578	0.6177	0.6184

Table 2: Comparison of the AUC for each classification algorithm.

The classifiers were compared graphically using ROC curves. The graph displayed in Figure 1 demonstrates that the tree-based classifiers were, as a group, superior to the other classifiers when compared using the AUC.

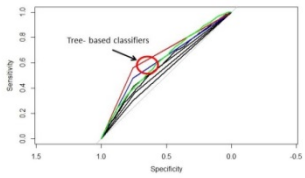

Figure 1: Comparison of classifiers using ROC curves.

The r program SuperLearner was used to compare the following classifiers LDA, QDA, GLM, random forest, boosting, and bagging (Table 3).

There was very little distinction amongst the models; but, for both the accuracy and the AUC, the ensemble classifier was the most accurate.

External validation was performed using a dataset from the author's practice. The practice maintains an independent database which contains the variables used to construct models from the large composite database.

However, for the external validation, a later time-period was selected.

The external validation set had 588 observations. Collins et al. have estimated that an external validation set needs to have over 200 observations to be valid [12], which was the case with this dataset.

The accuracy was compared to the accuracy of the original models and is summarized in Table 4.

Program	LDA	QDA	GLM	Boost	Bagging	Random Forest	Ensemble
Accuracies	0.7537	0.7131	0.7534	0.7438	0.7545	0.7441	0.7546
AUC	0.586	0.5776	0.5855	0.5887	0.5	0.589	0.6055

Table 3: Comparison of six classifiers used to create a new classifier.

Parameter	Linear classifiers					Tree-based Classifiers		
Program	GLM	LDA	QDA	KNN	SVM	TREE	FOREST	BOOST
Accuracy	0.7487	0.8062	0.7781	0.7368	0.7545	0.809	0.8689	0.8012
AUC	0.6425	0.6454	0.5784	0.5534	0.5147	0.6578	0.8615	0.6184

Table 4: Comparison of accuracy and AUC using the external validation dataset.

All models performed better using the external validation dataset. The random forest classifier performed markedly better than any other classifier. One explanation for the improved performance of the classifiers on the external validation data is that there is less noise in the dataset. The dataset is from a well-maintained database where the chance of error is small. Also, the group of patients tends to be more homogeneous than is evident for the larger, multicenter training set of data.

Discussion

The results of this study support the impression that big data can be of value in the clinical decision-making process. However, these results also emphasize that no single statistical approach provides maximum accuracy for all databases. This study substantiates the impression that different datasets will require investigation as to which algorithms are the most accurate for a particular set of data.

The process of IVF generates a vast amount of data. The premise is that the increased amount of data combined with advanced predictive analytics models will increase the accuracy of predictions. The data for this study was derived from an electronic medical record system that is utilized by over 140 IVF clinics in the United States. The dataset had over 88 variables and 135,000 cycles of IVF. Like most EMR data, manipulation was needed in order to provide usable data for the construction of models. While this reduced the number of observations and variables, there remained considerable data for model construction. Model formation involves using the majority of data to construct the model and then applying the model to a different dataset to determine the accuracy of model predictions. The approach

in this study used 80% of the data from the EMR to construct the models, and 20% as a test set to establish the accuracy of the model. The final assessment of accuracy involved an external dataset. The model created from the larger dataset was applied to a smaller dataset specific to one IVF center.

There are a number of ways to quantitate the ability of a model to predict the outcome variable. To determine the accuracy of the models, this study used both accuracies as determined by the r software and the AUC. These methods usually produce similar results, but based upon the dataset they may predict different predictive abilities [8]. The accuracy is based upon the distribution in the test dataset. If the dataset that is being used has a different class distribution, the accuracy may not be the same. Thus, depending upon the class distribution in the actual set of data being used for a given project, the AUC may be more useful when determining which classifier would provide the most accurate predictions. The outcome variable for this study is heavily biased to either 0 or 1 gestational sacs with there being far fewer twin gestations. As such, the AUC would be a more accurate way to distinguish the classifiers.

The ability to collect and use large amounts of data has spawned advances in the analytic tools used to examine the information held within such data. Methods such as K-nearest neighbors, support vector machines, and tree-based methods are available for data analysis. The increase in analysis options raises the question as to which methods will provide the most accurate and useful analytic approach. One way to address this question is to use ensemble learning. Ensemble learning assumes that using multiple different classifiers will identify the best classifier for both the given dataset and the circumstance being investigated. Ensemble learning can take on a number of forms. For this study, tree-based methods and stacking were used. In stacking, algorithms are trained and then these algorithms are combined to create a new composite algorithm. The idea underlying this approach is that every single algorithm might be a weak predictor; but by combining them, the new algorithm represents a much more accurate predictor.

Another type of ensemble learning method uses tree-based classifiers, an approach that has become popular in many analytic projects. Tree-based methods segment the predictor space into a number of simpler regions [6]. Once this division has been accomplished, the prediction is made from the mean or mode of the region where the training observation exists. Simple tree-based modeling frequently does not outperform supervised learning methods. Therefore, techniques have been developed to increase the accuracy of tree-based methods which rely upon producing many trees and then combining them into a single prediction. Two advantages of tree-based models include ease of interpretability and the ability to mimic how people think. These advantages might make tree-based methods more appealing and useful to clinicians as they consider whether to apply such models in practice. However, tree-based models may have relatively less predictive accuracy and be overly influenced by small changes in data. Boosting is a tree-based method that attempts to improve a model's accuracy. Elith et al. [13] describe boosting as a process based on the assumption that it is easier to find many lesser accurate models and combine them than it is to find one model fitted with high accuracy. The authors also distinguish bagging from boosting, noting that boosting is a sequential process. The boosting method grows trees sequentially, where each tree is grown based upon information from the previously grown tree. The underlying principle is that each model is grown on a modified set of derived variables, as opposed to the original data which could lead to overfitting. Boosting tends to form smaller trees which may prove helpful for data interpretation. Lastly, a random forest improves accuracy by decorrelating the trees[6].

Ensemble learning applied to tree-based classifiers outperformed linear regression and other classifiers. This is an important finding since the prediction classifiers being used today in ART are frequently linear regression-based. Such findings suggest that increases in available data allow better prediction models that can be utilized in a variety of classifiers to improve accuracy. These results also suggest that no one classifier can be applied to all datasets. The results from the large database identified the classification tree classifier as superior to the other classifiers. Tree-based classifiers provided a 12.5% increase in accuracy. The ROC curve demonstrated more significant improvement for all tree-based methods and for classification trees specifically. Finally, the AUC demonstrated a 13% improvement for the tree-based classifiers.

Models were derived from a large database consisting of numerous IVF programs. However, medical decision-making is local and ideally specific to the patient in question. The expectation for moving to a big data approach and more sophisticated analytic methods is that the accuracy of patient-specific predictions will be increased. A fortunate set of circumstances permitted the evaluation of a local, single-practice database. The results were illuminating and emphasized the utility of ensemble learning and big data to assist patients in their decision-making. Using the local database, the accuracy of the non-tree-based classifiers was increased by 37.4% and the tree-based accuracy was increased by 33%. The average accuracy of the non-tree-based classifiers was still less than the tree-based methods by 10.5%. Perhaps more importantly, the classifier that provided the best prediction for this dataset was not the classification tree; but rather, was the random forest classifier.

This study is limited by its lack of data from cycles where preimplantation karyotyping was done. The knowledge that an embryo is euploid may have a high correlation with the prediction in-question and thus may negate the advantage of big database analytics.

Conclusion

The results from this study demonstrate the usefulness of large datasets that are readily available to IVF practices and the multitude of classifier algorithms. The caveat is that no single algorithm will suffice for all datasets. Fortunately, the software has been developed that can be used without charge and that is easily customized for use by a single practice or physician. As the literature expands using more diverse analytic methods, individual clinicians can utilize these advances in their practice creating robust information for patients needing to make choices about their care.

The results of this study clearly demonstrate that the nature of medical information dictates that multiple classifiers be evaluated. The concepts central to ensemble learning provide a framework that can be successfully and easily utilized in ART.

References

1. SAS Insights (2018) Machine learning: what it is and why it matters.
2. Deo RC (2015) Machine learning in medicine. Circulation 132: 1920-1930.
3. SAS (2017) Statistics and machine learning at scale.

4. Zhou Z (2018) Ensemble learning.
5. Zhou Z (2012) Introduction. In Zhou Z, editor. Ensemble methods. Boca Raton: CR Press 1: 4.
6. James G, Witten D (2015) Classification. In James G, Witten D, Hastie T, Tibshirani R, editors. An introduction to statistical learning. New York: Springer Chapter 8: Page 303.
7. Dixit A (2017) Stacked generalization. In Dixit A, editor. Ensemble machine learning. Birmingham Chapter 8: page: 230.
8. StackExchange (2016) Cross-validated.
9. eIVF (2018) Data: Moving your fertility practice to the 21st century.
10. Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Stat Appl Genet Mol Bio 6: 25.
11. Coretes C, Vapnik V (1995) Support-vector networks. Mach Learn 20: 273-297.
12. Collins GS, Ogundimu EO, Altman DG (2015) Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. Stat Med 35: 214-226.
13. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Anim Ecol 77: 802-813.