Genetic algorithm—assisted machine learning for clinical pregnancy prediction in in vitro fertilization

Claudio Michael Louis, S Kom, B Eng; Nining Handayani, DVM, MBiomed; Tri Aprilliana, BPH; Arie A. Polim, MD, SpOG, DMAS, MBHRE; Arief Boediono, DVM, PhD; Ivan Sini, MD, FRANZOG, GDRM, MMIS, SpOG

BACKGROUND: A clinical pregnancy prediction model was developed by implementing machine learning technology that uses a combination of static images and medical data to calculate the outcome of an in vitro fertilization cycle.

OBJECTIVE: To provide a system that can accurately and sufficiently assist with decision making that is critical to in vitro fertilization cycles, primarily embryo selection.

STUDY DESIGN: Historical medical data, which consist of clinical information and a complete transferred embryo image dataset, of 697 patients who underwent unique in vitro fertilization were collected. Various techniques of machine learning were used, namely decision tree, random forest, and gradient boosting; each technique used the same data configuration for performance comparison and was subsequently optimized using genetic algorithm.

RESULTS: A prediction model with a peak accuracy of approximately 65% was achieved. Significant differences in the performances of the 3 selected algorithms were apparent. Nonetheless, additional metric measurements, such as receiver operating characteristic, area under the curve score, accuracy, and loss, suggested that the gradient boosting model performed the best in predicting clinical pregnancy.

CONCLUSION: This study served as a stepping stone toward the application of in vitro fertilization prediction models that use machine learning techniques. However, additional validation steps are required to boost the model’s performance for its implementation in the clinical setting.

Key words: clinical pregnancy, in vitro fertilization, machine learning, prediction model, single embryo transfer

From the IRSI Research and Training Centre, Jakarta, Indonesia (Mr Claudio, Mses Handayani and Aprilliana, and Drs Polim, Boediono, and Sini); Morula IVF, Jakarta Clinic, Jakarta, Indonesia (Ms Handayani and Drs Polim, Boediono, and Sini); Department of Obstetrics and Gynecology, School of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia (Dr Polim); Department of Anatomy, Physiology, and Pharmacology, Institut Pertanian Bogor University, Bogor, Indonesia (Dr Boediono).

Patient consent is not required because no personal information or detail is included in this study.

Cite this article as: Louis CM, Handayani N, Aprilliana T, et al. Genetic algorithm assisted machine learning for clinical pregnancy prediction in in vitro fertilization. Am J Obstet Gynecol Glob Rep 2022;XX:x.ex—x.ex.

Corresponding author: Claudio Michael Louis, S Kom, B Eng. michael.louis18@yahoo.com

© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

http://dx.doi.org/10.1016/j.xagrr.2022.100133

Introduction

The implementation of machine learning technology in in vitro fertilization (IVF) has become popular with various research studies on the topic geared toward improving the effectiveness of IVF or the patient experience. However, despite the extensive study, the existing clinical application of such technology is still rare. Multiple reasons that account for the lack of its applicability include the experimental nature and high-cost procurement of the technology. The former is reflected in most existing studies, whereas the latter refers to several readily commercialized artificial intelligence (AI)-powered automated embryo evaluation technology, such as the Vitrolife KID or iDA system. The Vitrolife system is among the most well-known and documented technology that has proven the feasibility of an AI-driven system in the IVF laboratory. Nonetheless, its use is restricted with the Vitrolife time-lapse incubator (EmbryoScope), which is costly and only accessible in a few IVF clinics.

This study was conducted to explore the possibility of developing an AI system that can improve the IVF treatment experience by providing patients with better information regarding the prognosis of their treatment, which is the likelihood of an embryo transfer resulting in pregnancy. The backend technology of this system would use an AI backend technology, which makes use of day 5 or 6 embryo images and patient’s clinical data during the treatment course, to predict the probability of a clinical pregnancy. The novelty of this research lies in pursuing an alternative technique that is more flexible and cost-effective to be adopted through the use of simpler methods and data that are more commonly available.

Materials and Methods

Method

Patients who underwent a single blastocyst transfer with complete clinical data were recruited. Static images and clinical data were used in a single prediction pipeline, a novelty of this study. Furthermore,
the image feature extraction, an essential step for converting the images into tabular data so that it could be combined with the patient clinical data, was performed fully automatically. To date, most research studies have focused on either using a single type of data or developing 2 separate models that perform independently using separate datasets. In contrast, this study aimed at creating a single prediction pipeline that directly combines the extracted image features and clinical data for model training and prediction tasks. The methods for image feature extraction and model development are discussed in the following section.

This section is separated into 2 subtopics, image extraction and prediction model. Although both functions are dependent on each other, they were established separately. Each section highlights the importance of each process to the research, its execution, and subsequent results.

**Image preprocessing**

Image preprocessing was performed before extraction to improve the quality of images for training. The original images contained noises, such as patches of images that were not relevant to the task at hand. To address this concern, we performed what is referred to as “image segmentation,”\(^1\)\(^2\) in which the specific areas in a given image are selected and “segmented” into different parts. The Canny edge detector,\(^3\) an edge detection operator that detects and highlights the edges, lines, and shapes of objects in an image, was used for the image segmentation using the OpenCV library,\(^4\) a Python library that specializes in functions of image manipulation, including computer vision and image processing.

**Image feature extraction**

To combine the images and tabular data for AI model training, standardization of those data into the same shape became necessary. In this particular case, the associated image data would have to be converted into tabular form.

The traditional technique of histogram of oriented gradients (HOG)\(^5\) was chosen for the feature extraction in this study to convert an image into a list of numerical data, which is achieved by way of a complex mathematical equation.

The HOG is implemented with scikit-image\(^6\) python library, which consists of various image-processing algorithms. However, the HOG would often result in tens of thousands of rows of data (depending on the size of the original image), presenting a potential complication. To alleviate this, we used the principal component analysis (PCA).\(^7\) The PCA is a dimensionality reduction method designed to transform a large dataset into a smaller dataset while preserving the same amount of information as much as possible. A larger dataset could cause various problems, including inefficient computation. Thus, having a concise dataset would be more appropriate. The PCA associates the input data by measuring the covariance to construct a new set of variables, the principal components. The resulting principal components become the base on which the new dataset would be created according to the selected relevant components as dictated by the PCA method configuration (mainly the desired size). In less technical terms, the PCA serves to minimize the size of a dataset while preserving the information in the original dataset as much as possible.

**Prediction model**

The core part of this article is in the development of the prediction model that was premised on existing techniques established by previous studies. The selection was based on the knowledge of the associated researchers, available resources, and, primarily, great relevance to the field.

Of note, 3 different machine learning techniques were chosen to pertain to their feasibility for the task at hand, decision tree (DT), random forest (RF), and gradient boosting (GB). Each technique is explained below.

**Decision tree.** The DT is one of the simplest but most effective machine learning classifications or prediction techniques. It operates a classification task based on a simple if/else algorithm (of a branching tree). The DT typically produces appropriate and satisfactory results, but its simplicity limits the range of tasks that can be performed. In addition, its classification algorithm provides transparency as opposed to a black-box type of algorithm, such as the Convolutional Neural Network (CNN). Furthermore, the DT generates a set of examinable rules that is easily understood if necessary.\(^8\)

**Random forest.** An improved adaptation of the DT, the RF is one of the algorithms referred to as an “ensemble algorithm,” which functions by combining various simple techniques into 1
unified prediction pipeline. In particular, it is a DT ensemble consisting of a group of weaker trees. The number of DTs in a single RF varies, but generally, there would be enough DTs to make the rule incomprehensible, making it closer to a black-box algorithm. Given its ensemble structure, the RF could often perform better than the DT at certain tasks, but not always.9

Gradient boosting. GB is an ensemble prediction model similar to the RF but with the implementation of “boosting.” In addition, unlike the RF, which is strictly an ensemble of the DT, GB could contain other types of weak learners, although the DT is notably more extensively used. The final output of a GB model is the weighted sum of several weaker learners. Nonetheless, boosting, as the core of GB, converts each weak learner into a more capable learner rather than merely using them to create a unified ensemble. GB performs the boosting function by making use of the gradient function.10

The 3 algorithms used the same dataset, each supported by its training setting and configuration, to identify which algorithm is most optimal for performing the task at hand. In addition, genetic algorithm (GA) is employed for model training with the most ideal parameter configuration to ensure the peak performance of each algorithm.

Genetic algorithm
GA11 is a supporting technique used in research to increase the effectiveness of prediction model training. It is referred to as an evolutionary algorithm, and it functions by repeating the training process of multiple model instances simultaneously to eventually find and/or create the best-performing model. Although success is not guaranteed (it is possible that among the hundreds of models trained, there will not be much improvement found), it is a much more effective and efficient way to use models than manually using different models one at a time.

Scoring metrics
As mentioned in the GA section, a fitness function is defined to gauge the performance of each model in a population by computing the balanced accuracy score of the model using the testing dataset. Moreover, other scoring metrics were measured to thoroughly examine and compare the performances of each algorithm:

• Accuracy is the measurement of how correct a model is in predicting the true value of each case. It is the simplest metric that could evaluate a prediction model’s performance but is susceptible to complications associated with data imbalance. Data bias could lead to a false high accuracy score as the model might be entirely incapable of recognizing a class with small sample data.
• Balanced accuracy12 is an “improved” version of accuracy that is designed especially for data with imbalanced class distribution, such as in this study. Instead of simply quantifying the ratio for a correct prediction, balanced accuracy is computed on the basis of 2 other metrics: sensitivity (true positive value) and specificity (true negative value). Balanced accuracy is defined as:

  \[
  \text{Sensitivity + specificity} \div 2
  \]

The equation measures the model’s predictive capability by calculating the combined rate of true values for both existing classes, which is less influenced by data imbalances.
• Loss is the measure of distance between the predicted result and the actual outcome, which is also defined as an error, the extent of detachment between the prediction results and reality. The specific loss function used in this study is called Brier loss,13 which calculates loss by using mean squared error. The lower the loss value, the more ideal the model is as it indicates that the prediction result is close to the actual outcome. In other words, the loss function measures how poorly the model performs.
• Area under the receiver operating characteristic (ROC) curve (AUC)14 score: The ROC is a plotted curve that illustrates a classifier’s performance and compares its sensitivity and specificity (as mentioned above). The AUC describes the score of the ROC curve. A high AUC score denotes a great capability of the model in making a correct prediction and vice versa.

All discussed metrics were calculated using separate testing data from that of training, to avoid any training influence on the resultant scores. Details of the
data usage are implied in the following section.

Materials

The data were obtained from Morula IVF Clinic, Jakarta, Indonesia. The selection criteria included patients who have completed their IVF treatments and underwent a single-embryo transfer. To maintain robust training data, patients whose records had missing clinical data and/or unsuitable embryo images were excluded from this study.

The medical data of patients were combined with the respective images of embryos for the training and prediction task. The medical data included the patient’s physical and clinical characteristics, such as age, body mass index, and drug use. Overall, data on 44 unique variables were collected (Supplementary 1). Initially, 1077 patients fulfilled the criteria, yet a total of 669 rows of data was established as the final training dataset. The sample size for the “not pregnant” and “pregnant” groups were 402 and 267, respectively. Only patients with embryo images of sufficient quality were included, and KNN imputation was performed for study subjects with a tolerable amount of missing medical data. K-nearest neighbors’ algorithm (KNN) imputation refers to a complex technique that fills the missing data by attempting to “predict” its values.

The static embryo image was converted into numerical features through the HOG before training, resulting in a massive full dataset size of $669 \times 41,474$. Moreover, the PCA was performed to reduce the image features’ dimensionality to $669 \times 64$, which represents the static transferred embryo image of each patient row. The $669 \times 64$ dataset was subsequently combined with the respective clinical data, resulting in a final data count of 669 rows $\times 124$ columns. The data were separated into 2 different sets for model training and testing, at a percentage ratio of 80:20, respectively.

Institutional review board approval

This study has been approved by the national institutional review board, the Ethics Committee of the Faculty of

| Experiment result |
|-------------------|
| Algorithm         | Balanced accuracy | Loss | AUC   | Accuracy |
| Decision tree     | 0.61              | 0.38 | 0.62  | 0.62     |
| Random forest     | 0.61              | 0.38 | 0.58  | 0.62     |
| Gradient boosting | 0.63              | 0.34 | 0.63  | 0.65     |

AUC, area under the receiver operating characteristic curve.

---

**FIGURE 2**
Random forest score per generation by genetic algorithm

**FIGURE 3**
Gradient boosting score per generation by genetic algorithm

**TABLE**
Experiment result

| Algorithm             | Balanced accuracy | Loss | AUC   | Accuracy |
|-----------------------|-------------------|------|-------|----------|
| Decision tree         | 0.61              | 0.38 | 0.62  | 0.62     |
| Random forest         | 0.61              | 0.38 | 0.58  | 0.62     |
| Gradient boosting     | 0.63              | 0.34 | 0.63  | 0.65     |

AUC, area under the receiver operating characteristic curve.

**Louis. Machine learning for clinical pregnancy prediction in in vitro fertilization. Am J Obstet Gynecol Glob Rep 2022.**
Medicine, University of Indonesia, according to the protection of human rights and welfare in medical research.

Result

The prediction models are displayed in Figures 1, 2, and 3. This research used GA to simultaneously perform training on multiple prediction models to simplify and accelerate the training process. Through GA, 1 best-performing model of each prediction algorithm was obtained (Table). The training process for each model was principally the same, using the same set of data but with different model parameters. In addition, Figures 1, 2, and 3 display the change in the accuracy of the models. The best-performing model was chosen, regardless of where it is situated in the timeline.

As listed, GB attained the best accuracy score of 0.65. Both the RF and DT obtained an identical score of 0.62 on account of the RT characteristic, which consisted of a combination of DTs with the outcome being the average value of those DTs. In addition to accuracy, GB also achieved better scores in balanced accuracy, loss, and AUC. Overall, this experiment demonstrated that GB is the best fit to perform the task of clinical pregnancy prediction using our particular dataset.

Discussion

This study demonstrated that the GB prediction model achieved the highest accuracy of 0.65. The results of this study have substantiated the potential approach of developing a model that is intended to predict clinical pregnancy in IVF. Reiterating the objective of this research, although the final results have yet to match the current state-of-the-art technology, we believe that the potential of applying AI in IVF with minimal resources and data has been sufficiently displayed. As is common with AI-based technology, more data of sufficient quality will directly improve AI performance. Nonetheless, this research has proven that even with a smaller dataset, it is possible to generate a simple yet practical prediction model for IVF-related outcomes.

Using relatively limited training data consisting of 669 rows, a model with a peak accuracy of 65% was achieved using a basic machine learning technique. VerMilyea et al.\(^\text{15}\) reported a final overall accuracy of 67.7%, although the nature of their research was eminently different from ours, in terms of the data, method, and goal. Varying degree of success is achieved in other studies; some studies obtained an accuracy score as high as 90%,\(^\text{16}\) and some studies obtained a similar rate of 60%,\(^\text{17}\) such as in this study and in the study of VerMilyea et al.\(^\text{15}\) Furthermore, the stages of each research vary greatly, with some still at the preliminary phase and others having attempted live commercialization of the technology.

The results of this study should serve as the base for more in-depth research on a prediction system in IVF, even with limitations in data availability. However, we have to emphasize that increasing the data size would be crucial depending on the intended goal of the research. For example, if research is conducted to release a commercialized product, increasing the amount of training data is advised to ensure the feasibility and reliability of the system.

Conclusion

Our study provided proof of the variability of machine learning in IVF practice as a decision support system. More training data and additional validation steps are required to boost model performance to be at an applicable level in the clinical setting.

Conflict of Interest

The authors report no conflict of interest.

REFERENCES

1. Haralick RM, Shapiro LG. Image segmentation techniques. Comput Gr Image Process 1985;29:100–32.
2. Pham DL, Xu C, Prince JL. Current methods in medical image segmentation. Annu Rev Biomed Eng 2000;2:315–37.
3. Xu Z, Baojie X, Guoxin W. Canny edge detection based on Open CV. IEEE ICEMI 2017: 53–6.
4. Bradski G. The OpenCV Library. Dr Dobb’s J Softw Tools 2000;120:122–5.
5. Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE Comput Soc Conf Comput Vis Pattern Recognit 2005;1:886–93.
6. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. PeerJ 2014;2:e453.
7. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab Syst 1987;2:37–52.
8. Steinberg D. CART: classification and regression trees. In: 2015. https://www.researchgate.net/publication/265031802_Chapter_10_CART_Classification_and_Regression_Trees
9. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. Int J Comput Sci Issues 2012;9:272–8.
10. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21.
11. Goldberg DE, Holland JH. Genetic algorithms and machine learning. Mach Learn 1988;3:95–9.
12. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. Int Conf Pattern Recognit 2010:3121–4.
13. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3.
14. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30:1145–59.
15. VerMilyea M, Hall JMM, Diakiv SM, et al. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Hum Reprod 2020;35:770–84.
16. Kutlu H, Avci E. Detection of cell division time and number of cell for in vitro Fertilized (IVF) embryos in time-lapse videos with deep learning techniques. Academia. 2019. Available at: https://www.academia.edu/41112951/Detection_of_Cell_Division_Time_and_Num ber_of_Cell_for_In_Vitro_Fertilized_IVF_Emb rys_in_Time-Lapse_Videos_with_Dee p_Learning_Techniques. Accessed 03 February 2020.
17. Passmore L, Goodside J, Hamel L, Gonzales L, Silberstein T, Trimmer J. Assessing decision tree models for clinical in-vitro fertilization data. 2003. Available at: http://citeseerx. ist.psu.edu/viewdoc/download?doi=10.1.1.131.7597 &rep=rep1&type=pdf. Accessed 18 June 2021.