ARTICLE

FERTILITY & REPRODUCTION

OPEN ACCESS

Machine Learning Approach to Predict Clinical Pregnancy Potential in Women Undergoing IVF Program

Nining Handayani[a,*1,2], Claudio Michael Louis[a,2], Alva Erwin[2,3], Tri Aprilliana[2], Arie A Polim[1,2,4], Batara Sirait[1,5], Arief Boediono[1,2,6], Ivan Sini[1,2]

[1]Morula IVF Jakarta Clinic, Jakarta, Indonesia
[2]IRSI Research and Training Centre, Jakarta, Indonesia
[3]Faculty of Engineering and Information Technology, Swiss German University, Tangerang, Indonesia
[4]Department of Obstetrics and Gynecology, School of Medicine and Health Sciences, Atmajaya Catholic University of Indonesia, Jakarta, Indonesia
[5]Department of Obstetrics and Gynaecology, Faculty of Medicine Universitas Kristen Indonesia, Jakarta, Indonesia
[6]Department of Anatomy, Physiology and Pharmacology, IPB University, Bogor, Indonesia

ABSTRACT

Objective: Hidden knowledge could be discovered within a large practical data of in vitro fertilization (IVF) practice. In this study, Machine learning–based data mining techniques were utilized to construct a reliable prediction model for clinical pregnancy in IVF.

Study Design: A retrospective cohort multicenter study involving 4.570 IVF cycles. All patients underwent fresh embryo transfer at either the cleavage or blastocyst stage between January 2015 and December 2019. The experiment focused on utilizing tree-based classifiers to generate and compare the most effective prediction model that could predict a clinical pregnancy through clinical data. Additionally, each classifier is optimized via a genetic algorithm technique, along with the selection of variables.

Results: Both the decision tree and random forest showed similar performance that was much better than the gradient boost. The two superior classifiers achieved a balanced accuracy of roughly 0.62. Additionally, each prediction model was shown to work optimally with different combinations of variables, with some variables being consistently included, such as female age, and some consistently excluded, which provides an insight into the relationship between the variables and each prediction model.

Conclusion: Machine learning algorithm remains effective for the purpose of data mining and knowledge extraction in IVF clinical datasets through which a relatively reliable prediction system for clinical pregnancy could be constructed, provided the available data is sufficient.

Keywords: In Vitro Fertilization; Prediction Model; Decision Tree; Machine Learning; Artificial Intelligence.

ABSTRAK

[Abstract in Bahasa Indonesia]

Tujuan: pengetahuan dapat diungkap dari data praktik fertilisasi in-vitro (FIV) dalam jumlah besar. Pada penelitian ini, teknik penambangan data berbasis pembelajaran mesin digunakan untuk membangun model prediksi kehamilan klinis pada program FIV.

Desain penelitian: multisenter kohort retrospektif menggunakan 4.570 siklus IVF. Seluruh subjek menjalani transfer embrio segar baik tahap *cleavage* maupun blastokista antara Januari 2015-Desember 2019. Eksperimen memanfaatkan pengklasifikasi berbasis *tree* untuk memperoleh model prediksi kehamilan klinis. Setiap pengklasifikasi dioptimalkan melalui teknik *genetic algorithm* bersama dengan seleksi variabel.

Hasil: *Decision tree* dan *random forest* mencapai kinerja prediksi yang lebih baik dibandingkan *gradient boost*. Kedua pengklasifikasi tersebut mencapai akurasi sekitar 0,62. Setiap model prediksi bekerja optimal dengan kombinasi variabel yang berbeda, dengan beberapa variabel digunakan secara konsisten, seperti usia perempuan, dan beberapa tidak digunakan secara konsisten. Hal tersebut memberi informasi tentang hubungan antara variabel dan setiap model prediksi.

[a]These authors contributed equally to this work.

Received 13 April 2022; Accepted 4 July 2022; Published

*Correspondence should be addressed to: Nining Handayani, DVM, M. Biomed, The BIC, Teuku Cik Ditiro street 12A, Menteng, Central Jakarta, Jakarta 13350, Indonesia. Email: nining.handayani@irsi-bunda.org

DOI: 10.1142/S2661318222500098

Kesimpulan: Algoritma pembelajaran mesin efektif untuk penambangan data dan ekstraksi pengetahuan dari data klinis FIV dalam jumlah besar. Model prediksi kehamilan klinis yang relatif andal dapat dikembangkan dengan ketersedian data.

Kata kunci: fertilisasi in-vitro, model prediksi, *decision tree*, pembelajaran mesin, kecerdasan buatan.

INTRODUCTION

Assisted reproductive technology (ART) has come a long way since its introduction over four decades ago, yet despite the continuous technological advances in the practices of in vitro fertilization (IVF), significant improvements in the success rate of IVF cycles have not been fulfilled. Overall clinical pregnancy rate remain low ranging from 26% to 36% (Andersen et al., 2007; de Mouzon et al., 2020; European Society of Human Reproduction and Embryology, 2018; Zegers-Hochschild et al., 2014). Several factors such as demographics, clinical characteristics, gametes quality, embryo quality, and endometrial aspects have been considered to influence the probability of pregnancy or live birth through an IVF program (Baker et al., 2010; Simopoulou et al., 2018; Nanni et al., 2010; Uyar et al., 2015; Vaegter et al., 2017). The complex association of these multivariable contributes to the difficulty of making a prediction. In general, IVF treatment requires multiple counseling sessions and multistage procedures, beginning with ovarian stimulation, ovum pick up (OPU), embryo culture, and subsequent embryo transfer procedure. Such complex procedures are costly, time-consuming, and somewhat stress-inducing to infertile couples. Therefore, an urgency to create predictive models that could forecast the outcome of IVF treatment has been required and persistently pursued by research attempts since 1989 (Hughes et al., 1989).

Numerous studies have been conducted to correlate the relationship between multiple IVF attributes and the different outcomes of IVF including clinical pregnancy, ongoing pregnancy, live birth, as well as cumulative live births (Ratna et al., 2020; van Loendersloot et al., 2010). Most studies have used statistical tools such as logistic regression and Cox regression due to the binary outcomes of interest. Women's age and duration of infertility were two variables that are used consistently to develop IVF prediction models followed by infertility causes, number of embryos transferred, number of previous IVF cycles, and embryo quality (Ratna et al., 2020). In another systematic review, it was found that pregnancy was negatively correlated with a female's age, duration of infertility, and basal FSH (follicle-stimulating hormone), while positively correlated with the number of oocytes retrieved (van Loendersloot et al., 2010). Each prediction model study seems to use different sample sizes and variables as predictors, which resulted in varying regression equations and models. Multiple articles and research exist that have discussed similar types of IVF outcome predictors including live birth (Barnett-Itzhaki et al., 2020; Ratna et al., 2020).

A growing number of research publications has shown enthusiasm for the utilization of artificial intelligence (AI) to improve the IVF success rate (Curchoe and Bormann, 2019; Simopoulou et al., 2018). The increasing trend is evident by the approximate seven-fold increase of AI-related manuscripts in the journals of human reproduction and embryology in 2018, which collectively suggested that AI could be used as leverage against current IVF complications (Curchoe and Bormann, 2019). The advantages of exploiting AI tools are not limited only to extracting information from a large and complex data set, but also to the processing of images, and transforming them into a desirable data level for further analysis. Nonetheless, IVF experts have attempted to utilize AI to automate several tasks such as cell counting (Khan et al., 2016) and embryo grading in the theoretical expectation that the automation could diminish the subjectivity of manual embryo observation, thus improving the consistency of embryo selection for transfer (Bormann et al., 2020).

Machine learning–based data mining approach is a subgroup of AI that is quite popular in the ART field (Hassan et al., 2020; Nanni et al., 2010; Passmore et al., 2003; Raef and Ferdousi, 2019; Vogiatzi et al., 2019). The definition of data mining could be explained as a method of extracting knowledge from a known set of data through various approaches, which include machine learning–based techniques (Durairaj and Ramasamy, 2016; Nanni et al., 2010). Utilizing a large existing IVF clinical data, machine learning approaches allow the extraction and identification of hidden knowledge or unknown interrelationships between a number of IVF attributes and clinical pregnancy events, which in turn could be used to generate a reliable pregnancy prediction model.

Such a prediction model can be generated through many different algorithms. A few of these are commonly used for IVF-related predictions, namely tree-based classifiers and neural networks (including their many variants and derivatives) (Louis et al., 2021; Raef and Ferdousi, 2019). Every existing prediction model carries its own characteristics, and each is suited to a different type of task. Neural network, one of the most commonly used techniques in present times, presents itself as one of the most effective methods of prediction for all types of data including visual and tabular data. It, however, offers a weakness in terms of expensive computing cost and a black-box type of operation, which means details on the operation is relatively scarce (Du and Swamy, 2014). While the neural network still proves the best solution for visual data such as images and videos, in terms of tabular data, tree-based classifier offers an equally effective solution for a much cheaper computing cost and clarity in terms of process detail. Consequently, this study was conducted to develop tree-based prediction model from clinical data that can be utilized to predict the occurrence of clinical pregnancy in IVF patients. To achieve this goal, we concurrently sought predictive variables that could establish the most optimum prediction model.

MATERIALS AND METHODS

Subject characteristics

This multicenter retrospective cohort study evaluated IVF clinical pregnancy data obtained from three Morula IVF Indonesia clinics located in Jakarta, Surabaya, and Makassar during the period between January 2015 and December 2019. A total of 4,570 cycles were included. The information collected encompassed baseline and clinical characteristics of study subjects, embryology laboratory results, and clinical pregnancy outcomes. All women included in this study underwent fresh embryo transfer at either cleavage stage (day 2/3) or blastocyst stage (day 5/6). A total of 4,570 labeled data were retrieved from the online databases of the three clinics. The data were divided into an 80/20 split for training/testing. The 80% of data are utilized entirely for training, which involve the process of hyperparameter tuning and feature selection that was achieved by utilizing genetic algorithm (GA). A 10-fold cross-validation is used in both processes to ensure an effective training result. The remaining 20% is reserved until the end of both hyperparameter tuning and feature selection, and is used only to gauge the performance of the final chosen model. This practice is a common method to evaluate a prediction's model performance during training because by

Table 1. List of variables used in this study.

No	Variable	Nature of data	Min-Max
1	Female's age (years)	Numeric	23 – 47
2	Stimulation method	Nominal	N/A
3	Number of previously failed IVF treatment(s)	Numeric	1 – 7
4	Type of infertility	Nominal	N/A
5	Duration of infertility (years)	Numeric	0.5 – 20
6	History of miscarriage	Category	N/A
7	Female BMI (kg/m^2)	Numeric	15.98 – 40.51
8	IVF indication: Endometrial factor	Category	N/A
9	IVF indication: Sperm factor	Category	N/A
10	IVF indication: Recurrent IUI failure	Category	N/A
11	IVF indication: Unexplained factor	Category	N/A
12	IVF indication: Other factors	Category	N/A
13	Female prognosis	Nominal	N/A
14	Basal FSH (mIU/mL)	Numeric	1.81 – 20.80
15	Basal LH (mIU/mL)	Numeric	0.20 – 17.30
16	Basal estradiol (E2) (pg/mL)	Numeric	5 – 381
17	Basal progesterone (P4) (ng/mL)	Numeric	0.05 – 6.74
18	AMH (ng/mL)	Numeric	0.11 – 21.33
19	AFC	Numeric	1 – 30
20	Estradiol level on trigger day (pg/mL)	Numeric	207 – 7249
21	Progesterone level on trigger day (ng/mL)	Numeric	0.05 – 4.51
22	Type of gonadotropin	Nominal	N/A
23	Starting dose of gonadotropin	Numeric	75 – 375 IU
24	Type of suppression drug	Nominal	N/A
25	Type of maturation trigger drugs	Nominal	N/A
26	Number of oocyte(s) retrieved	Numeric	1 – 53
27	Number of mature oocyte(s) following injection	Numeric	1 – 42
28	Maturation rate (%)	Numeric	12 – 100
29	Sperm quality	Nominal	N/A
30	Number of fertilization(s)	Numeric	1 – 35
31	Number of cleavage(s)	Numeric	1 – 33
32	Number of top-quality cleavage(s)	Numeric	0 – 20
33	Number of blastocyst(s)	Numeric	0 – 26
34	Number of top-quality blastocyst(s)	Numeric	0 – 15
35	Day of embryo transfer	Nominal	N/A
36	Number of top-quality ET(s)	Numeric	0 – 3
37	Total number of embryo(s) transferred	Numeric	1 – 3
38	All top-quality ET	Category	N/A
39	Mix quality ET	Category	N/A
40	Female smoking status	Category	N/A
41	Male smoking status	Category	N/A
42	Male alcohol drinking history	Category	N/A

AFC, antral follicle count; AMH, anti-Müllerian hormone; BMI, body mass index; ET, embryo transfer; FSH, recombinant follicle-stimulating hormone; IUI, intra-uterine insemination; IVF, in vitro fertilization; LH, Luteinizing hormone.

maintaining a separate dataset, a nonbiased scoring of the prediction results is guaranteed. The percentage of the split may vary but is usually around 20% or 30% for the testing portion (Hassan et al., 2020). About 20% was decided arbitrarily, as it is also the most common amount when it comes to splitting a dataset. The data set comprised 1,669 clinically pregnant women and 2,901 nonpregnant women. Of all the 70 recorded variables, 42 attributes were selected as potential predictors according to their significance in influencing clinical pregnancy events based on existing literature, data availability, and clinical experts' opinions in our IVF clinics (Table 1).

Data set preparation

Preprocessing of raw data was conducted using data classification tools in Microsoft Excel before imputation into the machine learning application. Missing values were sought by tracking the hard copy files to ensure complete data sets. Subjects whose data were missing for several important attributes (body mass index [BMI], infertility duration, and complete loss of basal hormonal result) were excluded.

Tree-based classifiers

A tree-based classifier is one of the earliest and most commonly used types of classifier for the prediction tasks due to its effectiveness and cost-efficiency, which have been adopted for various functions including in IVF (Louis et al., 2021; Raef and Ferdousi, 2019). Compared to a more modern technique, tree-based classifiers require a relatively low computing cost, while still yielding an adequate or even good performance. The basis of tree-based classifiers is akin to creating a set of rules that determines the value of a target attribute depending on a set of if-else conditions. This method is decidedly quite simple but has been proven effective for a lot of different cases (Charbuty and Abdulazeez, 2021). An additional advantage is the resulting set of rules that is comprehensible, as opposed to the total black-box function of other more advanced prediction model algorithms.

In this study, we selected three types of tree-based classifiers, namely decision tree (DT), random forest (RF), and gradient boosting (GB). RF and GB are both derivative algorithms, called ensemble methods that are based on DT. The ensemble method refers to an algorithm that works by combining other simpler algorithms (such as DT) into one giant collective function. In this case, both RF and GB are an ensemble method of DT, each performing its ensemble process in a different way.

Genetic algorithm

In the process of creating a prediction model, a concept called hyperparameters is critical to the results. These hyperparameters refer to a set of attributes that are associated with the chosen algorithm and affect the algorithm's behavior according to their values. The configuration of these hyperparameters, also known as hyperparameter tuning, is one of the primary parts of prediction model development. This tuning could be done manually, which could be proven to be inefficient as it requires individual experimental testing of every possible combination of the variables. Thus, a solution called optimization algorithm, for instance GA, has been created to automate the process of hyperparameter tuning.

GA is an evolutionary algorithm that is based on the concept of biological evolution (Goldberg and Holland, 1988). This algorithm mimics the evolving process of biological lifeforms such as reproduction to generally get better or become something more with each generation. The function of GA is centered around its population, which consists of individuals who possess possible solutions to the problem that GA is trying to solve. Each individual's potential for solving the target problem is measured through a fitness function. This population will then be optimized for maximum performance through the aid of operators such as selection, crossover, and mutation. Selection keeps and elects the best-performing individual to be utilized as a "parent," crossover creates new individuals by generating offspring from the chosen parents, and mutation mimics the random spontaneous changes in biological lifeforms, as a measure to maintain diversity in the population. Eventually, the main goal of the operators is to identify the best individual as the optimal solution for the problem at hand.

This technique is implemented in our research for hyperparameter tuning to optimize the set of hyperparameters for each of our chosen algorithms. This is achieved by creating a population of models of the chosen algorithm, each with different parameters combination, which is then repeatedly trained and contested through GA to achieve the best performing model as listed in Prediction model training. This approach is very similar to the one that was utilized by (Guh et al., 2011).

Features selection and data preprocessing

In developing a prediction model, the data utilized are as important as the method. The entire concept of data mining is to extract information from rows of data, so a poor-quality dataset would be reflected in an incapable prediction model. It is therefore important to perform an extensive preprocessing of data, which includes the preliminary data cleaning. Before the experiment, data cleanup was conducted.

Data processing procedure, in the development of a prediction model, is unique to data mining, in terms of preparing a dataset for training and testing (Alasadi and Bhaya, 2017; Kotsiantis and Kanellopoulos, 2006). Feature selection is one of the very straightforward procedures for data preparation. The process reduces the dimensionality of the dataset column-wise, by discarding the least influential attributes, thus leaving only those that are expected or are proven to have a significant influence in predicting the target outcome. GA could be utilized for the feature selection function. In a population, each individual contains a different combination of variables, and over several generations, the best-performing individual who is equipped with variables that produce a superior prediction model could be achieved.

Aside from feature selection, additional data preprocessing methods, namely missing value imputation and variable encoding, were introduced to the dataset before training. As discussed previously, the major missing values were handled manually in Microsoft Excel with some minor fields that were still left empty. The amount of missing data varies from each row, and while not all row contains missing fields, a handful was noted to be missing, with the percentage ranging from 2% to 11%, for each row. Simply dismissing every single row with a missing field might incur a lot of knowledge loss, therefore, a method called imputation was applied to prevent it. Imputation here refers to a process of filling the empty fields of data, either through a simple method of using the attribute average, or a complex calculation approach. On account of the non-normal data distribution, KNN (K-Nearest Neighbor) imputation was applied to resolve the remaining missing values in the data set.

KNN is a machine learning algorithm that was adapted for imputation. The basis of this algorithm revolves in finding the similarity between data points and subsequently performing either imputation or classification based on the similarity. In the case of imputation, a KNN algorithm determines the value of the missing variables by comparing the distance between the rows of missing values to every other row in the dataset, with nonmissing variables (Zhang, 2012).

Development of prediction model in python

The experiment for this research was carried out entirely in the Python programming language, which utilizes many different existing laboratories that could be suited especially for the machine learning tasks (Buitinck et al., 2013). Scikit-learn is a prominent and commonly adopted machine learning library for python that provides many ready-to-use algorithms for tasks such as classification and regression (Buitinck et al., 2013).

In terms of DT, which have many iterations, scikit-learn employs a variant of the Classification and Regression Tree (CART) (Steinberg, 2009). As both GB and RF are extensions of DT, all three methods technically make use of the same base algorithm in CART but each in its own way. The GA in this research is developed from zero, and thus is a custom implementation in a sense.

Metrics of evaluation and their definitions

The capability of a prediction model can be examined through various metrics, which provide different insights on the model's performance as each would highlight a different area (Raef and Ferdousi, 2019). Each metrics evaluates different aspects of a model, and their significance depends on the objective and method of the research or application that is being carried out. Out of all existing metrics, we choose accuracy for its straightforward nature, which simply measures the frequency of correct predictions made by the model against the existing total example. Mathematically, accuracy can be defined as, with TP referring to true positive, TN as true negative, FP as false positive, and FN as false negative.

$$\text{(Accuracy)} \qquad \frac{TP+TN}{TP+TN+FP+FN}$$

Weakness exists, however, in utilizing accuracy as a metric, primarily in how it is dependent on prevalence, vulnerability to data imbalance. Hence, variations of the accuracy metric were developed to combat its weaknesses. One such variation is referred to as balanced accuracy and is a variation of the accuracy metric that is designed to deal with and take into account class imbalance in a dataset. It is defined as "the average of recall obtained on each class." Recall itself is a measure of the prediction model to find positive samples, and is measured for each class. Balanced accuracy essentially averages this value for every class to measure the prediction model's capability. This allows for an accuracy metric that is not affected by data imbalance and instead is measured depending on the portion of correct prediction that was made for each class. In our case, balanced accuracy can be mathematically described as:

$$\text{(Balanced Accuracy)} \qquad \frac{1}{2}\left(\frac{TP}{FP+FN}+\frac{TN}{TN+FP}\right)$$

Aside from accuracy and balanced accuracy, three other metrics were examined, namely precision, recall, and F1-score. Recall was already calculated as part of balanced accuracy, but here it is also examined separately. The scores of the three metric measurements were separately calculated for each class. The metrics can be defined as:

$$\text{(Precision)} \qquad \frac{TP}{TP+FP}$$

$$\text{(Recall)} \qquad \frac{TP}{TP+FN}$$

$$\text{(F1 Score)} \qquad \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

Precision describes the model's ability to correctly label positive samples against negative samples. Recall describes the model's ability to find all positive samples, as mentioned earlier, and F1 score is a weighted mean of both precision and recall in which a score of "1" denotes a perfect record. These additional metrics could add more insight into the model's performance, preventing its misjudgment that could arise from contemplating only at the accuracy score.

RESULTS

Prediction model training

The overall clinical pregnancy rate of our data set was 36.5% (1.669/4.570). The results of the three selected algorithms were presented in Table 2. Due to the automatic nature of GA, it is thus not possible to list the results of every parameter and feature combination, as there are thousands of combinations that were attempted in total.

As discussed previously, accuracy here denotes the overall predictive ability of the model, as in how many predictions are correct overall. The other three scores are separated into two, one for each class, which denote the metric measurement for the relevant class. Precision score for the label (0) denotes how many nonpregnant cases were correctly labeled, in respect to the entire population that was labeled as nonpregnant. The precision for DT, for example, denotes that from all of the test cases labeled as not pregnant, 80% of them are correct with the remaining being false positives, that is, pregnant cases labeled as not pregnant. Recall describes the model's ability to find the entire population of a certain label. For example, the DT was able to correctly find approximately 46% of the not-pregnant case, and the remaining 54% being misclassified as another class (in this case, as pregnant). F1 score, as mentioned before, is simply a weighted mean of both precision and recall, which is of course also calculated for each label.

The change in precision and recall for each model through the process of hyperparameter tuning and feature selection were depicted in Figures 1, 2, and 3. Generally, the GA implementation is shown to be effective in maximizing the performance of the three chosen models through parameter tuning, and has a consistent pattern of dropping before eventually climbing up and plateauing at a certain value (which differs for each model). Feature selection however achieves a more inconsistent result, and although some

Table 2. Summary of model evaluation metrics.

		Precision (%)		Recall (%)		F1 Score	
Model name	Accuracy (%)	0	1	0	1	0	1
Decision Tree	0.63	0.80	0.44	0.46	0.79	0.59	0.56
Random Forest	0.61	0.75	0.45	0.59	0.63	0.66	0.53
Gradient Boost	0.58	0.69	0.51	0.83	0.32	0.76	0.40

Note: (0) denotes for row labeled as nonpregnant and (1) for pregnant. AUC, area under the curve.

Fig. 1. Result of genetic algorithm for decision tree (A) hyperparameter tuning (B) feature selection.

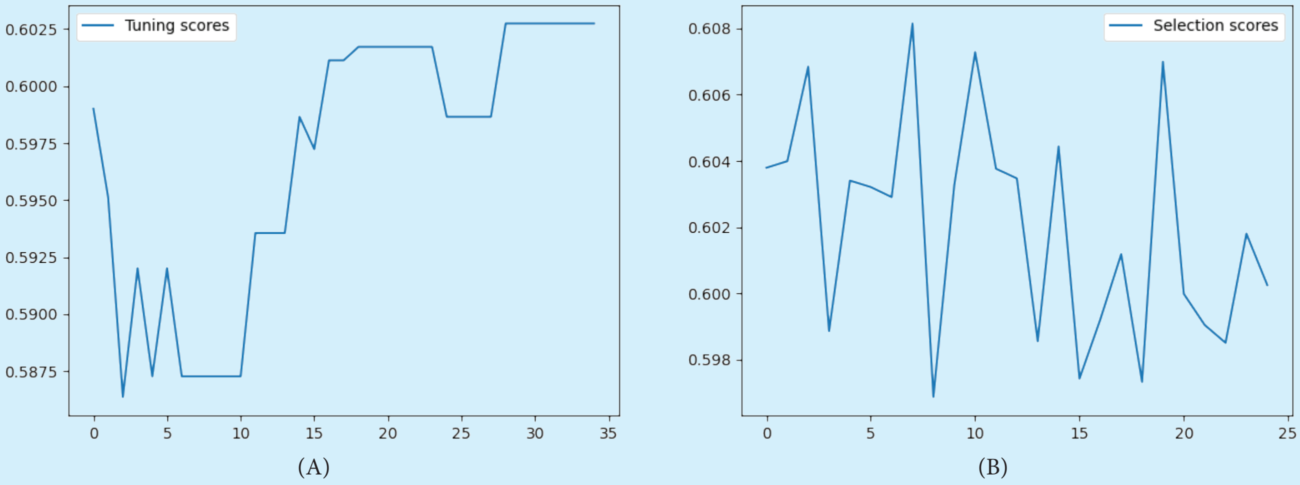

Fig. 2. Result of genetic algorithm for random forest (A) hyperparameter tuning (B) feature selection.

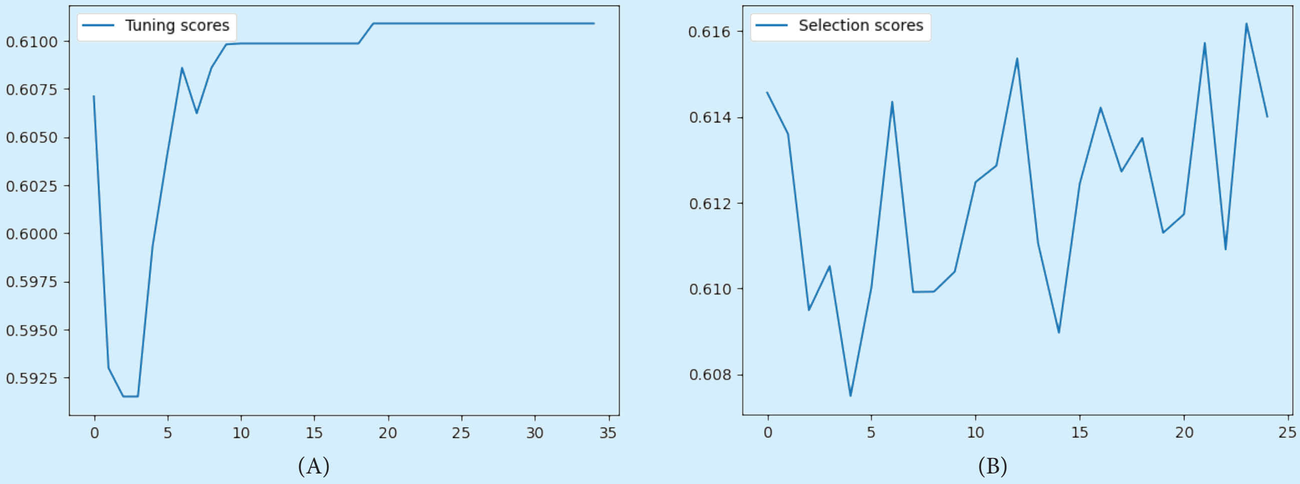

Fig. 3. Result of genetic algorithm for gradient boost (A) hyperparameter tuning (B) feature selection.

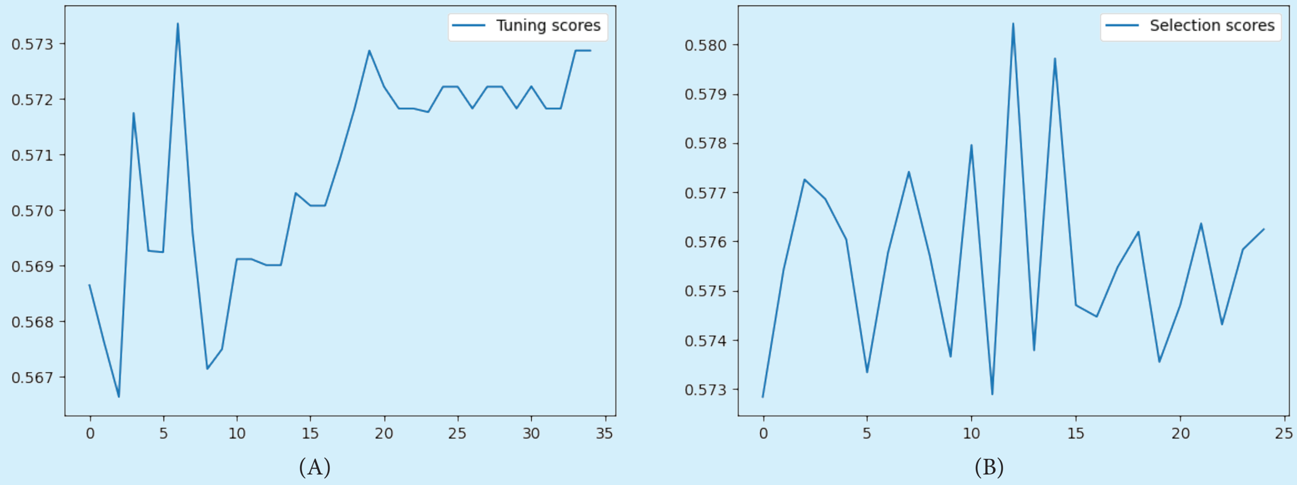

Table 3. List of chosen variables for each model.

	Model		
Features	**Decision tree**	**Random forest**	**Gradient boost**
Female's age (years)	√	√	√
Stimulation method	-	-	-
Number of previously failed IVF treatment(s)	√	√	√
Type of infertility	-	√	√
Duration of infertility (years)	-	√	-
History of miscarriage	-	√	√
Female BMI (kg/m^2)	√	√	√
IVF indication: Endometrial factor	√	-	√
IVF indication: Sperm factor	√	-	√
IVF indication: Recurrent IUI failure	√	√	√
IVF indication: Unexplained factor	-	-	√
IVF indication: Other factors	√	√	√
Female Prognosis	-	√	√
Basal FSH (mIU/mL)	√	-	-
Basal LH (mIU/mL)	√	-	-
Basal estradiol (pg/mL)	√	√	√
Basal progesterone (ng/mL)	-	√	√
AMH (ng/mL)	-	-	√
AFC	-	-	-
Estradiol level on trigger day (pg/mL)	-	-	-
Progesterone level on trigger day (ng/mL)	-	√	√
Type of gonadotropin	√	√	√
Starting dose of gonadotropin (IU)	√	√	-
Type of suppression drug	√	√	√
Type of maturation trigger drugs	-	√	√
Number of retrieved oocytes	-	√	√
Number of mature oocytes following injection	√	-	-
Maturation rate (%)	-	-	√
Sperm quality	√	√	√
Number of fertilization(s)	-	√	-
Number of cleavage(s)	-	√	√
Number of top-quality cleavage(s)	√	√	√
Number of blastocyst(s)	√	√	-
Number of top-quality blastocyst(s)	√	√	√
Day of embryo transfer	√	√	√
Number of top-quality ET	√	√	√
Total number of embryo(s) transferred	√	-	-
All top-quality ET	-	√	√
Mix quality ET	-	√	√
Female smoking status	√	-	-
Male smoking status	√	√	-
Male alcohol drinking history	-	-	√

generations were able to reach a peak value, it eventually drops and rises again at a random interval.

Each of the scores in Table 4 is a result of both hyperparameter tuning followed by feature selection, both performed through the variant of GA designed specifically for each task, measured with the metrics as explained in the earlier section. Each model is processed through GA with a population of 50 on 35 generation for hyperparameter tuning, and with a size of 100 for 25 generation for feature selection. This does mean that the hyperparameter tuning is done on the full attribute, thus not taking into account the possible

Table 4. Feature importance for each model.

	Model		
Features	**Decision tree**	**Random forest**	**Gradient boost**
Female's age (years)	0.30830732	0.30097652	0.16517494
Stimulation method	-	-	-
Number of previously failed IVF treatment(s)	0	0.00227324	0
Type of infertility	-	0.00393157	0
Duration of infertility (years)	-	0.03431229	-
History of miscarriage	-	0.00144527	0
Female BMI (kg/m^2)	0	0.06621351	0.07594501
IVF indication: Endometrial factor	0	-	0
IVF indication: Sperm factor	0.00159487	-	0
IVF indication: Recurrent IUI failure	0	0.00112629	0
IVF indication: Unexplained factor	-	-	0
IVF indication: Other factors	0	0.00627505	0
Female prognosis	-	0.00341621	0.00295585
Basal FSH (mIU/mL)	0	-	-
Basal LH (mIU/mL)	0	-	-
Basal estradiol (pg/mL)	0.04939534	0.04166176	0.08605917
Basal progesterone (ng/mL)	-	0.05050891	0.07816868
Anti-Müllerian hormone (AMH)	-	-	0.08527922
Antral follicle count (AFC)	-	-	-
Estradiol level on trigger day (pg/mL)	-	-	-
Progesterone level on trigger day (ng/mL)	-	0.12213946	0.13493776
Type of gonadotropin	0.01950395	0.00311776	0.018364805
Starting dose of gonadotropin (IU)	0.04532669	0.01394578	-
Type of suppression drug	0	0	0
Type of maturation trigger drugs	-	0.00281226	0
Number of retrieved oocytes	-	0.0323469	0.04268476
Number of mature oocytes following injection	0	-	-
Maturation rate (%)	-	-	0.04285337
Sperm quality	0	0.00526807	0
Number of fertilization(s)	-	0.02317381	-
Number of cleavage(s)	-	0.01358787	0.08325507
Number of top-quality cleavage(s)	0	0.019504	0.0043817
Number of blastocyst(s)	0	0.03236228	-
Number of top-quality blastocyst(s)	0.09437997	0.03237783	0.03736177
Day of embryo transfer	0	0	0
Number of top-quality ET	0.41999114	0.17514111	0.0473779
Total number of embryo(s) transferred	0.06150071	-	-
All top-quality ET	-	0.00504135	0.05274042
Mix quality ET	-	0.0024206	0.02409477
Female smoking status	0	-	-
Male smoking status	0	0.00462033	-
Male alcohol drinking history	-	-	0

selection of attributes. A more complete method would perhaps to perform both tuning and selection simultaneously, but such process will require a very expensive computing cost. Therefore, the alternative of performing tuning on a full data set followed by selection was chosen. This decision was made after observing that, with the dataset provided, the process of feature selection does not offer much change in performance when done on an unoptimized model.

Selection of predictive variables

Since feature selection was performed individually for each classifier, in total, three sets of features have been selected specifically per classifier. The selection method is intuitive through GAs by which a feature set unique to each model that produces the best accuracy will be used. While each model had different optimal feature sets, certain patterns and similarities could be observed which indicate the importance of some features. Considering all three models are based on a DT, determining the relation of the chosen variables with the tree-based classifiers is possible. For example, in all models, female age was chosen as one of the variables, which displayed its prominent influence in performing pregnancy prediction. Concurrently, variables that exhibited the least predictive potential (those that were consistently excluded) could also be identified in the results of our dataset (Table 3).

Moreover, to expand from merely just choosing between the various features, we also measured the importance of the features for each model (Table 4). The list of feature importance highlights the fact that, although DT "choose" a lot of features, due to its simplistic nature, only a few of the chosen features were actually utilized to perform the expected prediction. We can also see that the initial expectation of female age being the most important is almost correct, as it is shown to achieve the highest importance score for two of the three models. For DT model, while female age achieves the second highest importance, we observed that the number of top-quality embryo transfer were given a higher importance by a considerable margin, while having a more average importance for both Random Forest and Gradient Boost.

DISCUSSION

This study has displayed the benefits of utilizing machine learning–based data mining concepts to derive knowledge from the considerably large IVF clinical database retrospectively. This study has shown that DT and RF achieved a comparable prediction performance with a balanced accuracy of roughly 0.62 as compared to gradient boost which only achieved a balanced accuracy score of 0.58. DT algorithm is widely used to classify binary outcomes and presents an intuitive set of rules. The optimal combination of variables for each predictor model was established through GA, and while the sets were different, certain variables were consistently included in all models such as female age, number of top-quality blastocyst(s) and number of top-quality ET. This has corroborated previous studies and served as proof, which implies the significance of female age in predicting pregnancy.

In the present time, classical statistics and machine learning are common complementing tools for the construction of a mathematical model to predict IVF outcomes. However, it has been suggested that machine learning algorithm such as neural network is more powerful in recognizing a broad-range nonlinear association among variables compared to other statistics analysis (Kaufmann et al., 1997). Additionally, a prediction model constructed through machine learning algorithms offers a sufficiently solid forecasting accuracy over the K fold cross-validation method. The potential advantages of machine learning algorithms over classical statistics in attaining true IVF outcome predictions have also been highlighted in a recent study (Barnett-Itzhaki et al., 2020).

Generally, our models displayed sufficient accuracy but not as high as the existing studies (Guh et al., 2011; Hafiz et al., 2017; Hassan et al., 2020). Differing results among studies, including ours, could be attributed by the different data sets and evaluation metrics that were used in each study; thus, comparing the result would be inappropriate. In addition, sample sizes and predictor variables that were introduced to the machine learning algorithms were slightly different. To date, there has yet to exist a consensus protocol that could specify which variables hold the most predictive potential in characterizing a successful IVF cycle (van Loendersloot et al., 2010). This issue is of most importance in machine learning studies because different data sets and diverse variables would produce different sets of rules and prediction models that are unique to each respective data set.

Since the balanced accuracy of the derived models did not achieve at least 0.7, we considered three points of discussion in elucidating these results. First, the noninclusion of important endometrial aspects may carry significant values in distinguishing between pregnant and nonpregnant cases. For instance, the endometrial thickness variable was excluded due to a substantial proportion of unrecorded data before 2017 in our IVF Centre's database. The essential role of endometrial aspects in predicting pregnancy through machine learning algorithms has also been demonstrated (Nanni et al., 2010). In the preliminary study, Nanni et al. attained an area under the curve (AUC) of up to 0.85 by employing sub-endometrial volume, endometrial vascularization index, or flow index combined with female age.

Second, we might not be able to achieve high accuracy when subjects with unspecified infertility were used to predict clinical pregnancy due to the complexity of the multivariable data. In this case, we propose to build a prediction model based on the similarities of the study subjects (e.g., high, normal, and poor responder group) or based on the infertility causes. Another approach is to create the prediction model based on each procedure of the IVF treatment such as an ovarian response prediction model, oocyte retrieval prediction model, blastocyst prediction model, and/or a live birth prediction model. Consequently, noise and vagueness in the relationship between independent and dependent variables might be minimized. Third, the heterogeneity origin of the multicenter-derived data might influence the results of this study.

Certain measures can be taken to solve such issues, for instance by conducting a more selective data selection or introducing new variables that were previously excluded. Referring to existing research to utilize the same set of variables is speculated to offer similar (or better) results, yet the variety of the patient records and the experiment execution to achieve certainty are to be considered. Pregnancy complications are often unexpected and are caused by unexplained factors that were reflected in our dataset. This ultimately challenged the reliability of the prediction model.

Our results possess beneficial inputs for developing countries such as Indonesia. As government subsidies for IVF treatments are nonexistent, creating the prediction model becomes essential for both the patients and clinicians. From an infertile couple's point of view, the availability of such a prediction model could aid in defining a rational expectation of their IVF success rate. Likewise, clinicians could benefit from the calculation to prevent the prescription of unnecessary and overused treatments. Locally, the important findings of this study have presented a novel opportunity to create a supportive decision-making system for our IVF centers.

The strength of this study is reflected in the use of the large, reliable retrospective IVF database to construct the prediction model. The main limitation of this study was the nature of the retrospective data collection. Further research worth investigating would be to evaluate other independent variables that could define the pregnancy outcomes more accurately. Otherwise, developing prediction models appertaining to each step of the IVF treatment may enhance the predictive performance.

CONCLUSION

The utilization of machine learning algorithms has permitted the extraction of statistical knowledge patterns from a large IVF database. Through the result of our research, both the DT and RF algorithms have a potential in being utilized to build an effective

AI-based clinical pregnancy prediction for use in IVF treatment. Each algorithm achieves a similar result, with DT having marginally higher score than RF through the metrics (accuracy, precision, recall, and the F1 score).

CONFLICT OF INTEREST

The authors declared that they have no conflict of interest or competing interest to disclose.

AVAILABILITY OF DATA AND MATERIAL

Data are available upon reasonable request.

ETHICAL APPROVAL

This study was performed in line with the principles of the Declaration of Helsinki. The Ethics Committee of Faculty of Medicine, University of Indonesia, have approved the study protocol.

INFORMED CONSENT

Because of the nature of retrospective study, waiver of informed consent was granted by the Ethics Committee of Faculty of Medicine, University of Indonesia, on July 6, 2020.

AUTHOR'S CONTRIBUTIONS

The original concept was designed by NH and MCL. NH and TA performed data collection and statistical analysis. MCL performed and created a machine learning model. AE supervised machine learning trial. NH, MCL, and AE drafted the manuscript. The manuscript was carefully corrected by AAP, BIS, AB, and IS. All authors have accepted the final version of the submitted manuscript.

REFERENCES

Alasadi SA, Bhaya WS. Review of data preprocessing techniques.pdf. J Eng Appl Sci. 2017;12(16):4102–7.

Andersen A, Goossens V, Gianaroli L, Felberbaum R, De Mouzon J, Nygren KG. Assisted reproductive technology in Europe, 2003: results generated from European registers by ESHRE. Hum Reprod. 2007;22(6):1513–25. https://doi.org/10.1093/humrep/des255

Baker VL, Luke B, Brown MB, et al. Multivariate analysis of factors affecting probability of pregnancy and live birth with in vitro fertilization: an analysis of the society for assisted reproductive technology clinic outcomes reporting system. Fertil Steril. 2010;94(4):1410–6. https://doi.org/10.1016/j.fertnstert.2009.07.986

Barnett-Itzhaki Z, Elbaz M, Butterman R, et al. Machine learning vs. classic statistics for the prediction of IVF outcomes. J Assist Reprod Genet. 2020;37(10):2405–12. https://doi.org/10.1007/s10815-020-01908-1

Bormann CL, Thirumalaraju P, Kanakasabapathy MK, et al. Consistency and objectivity of automated embryo assessments using deep neural networks. Fertil Steril. 2020;113(4):781–7. e1. https://doi.org/10.1016/j.fertnstert.2019.12.004

Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project. ArXiv:1309.0238.2013:1–15. http://arxiv.org/abs/1309.0238

Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends. 2021;2(01):20–8. https://doi.org/10.38094/jastt20165

Curchoe C, Bormann C. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. J Assist Reprod Genet. 2019;36:591–600. https://doi.org/10.1007/978-3-030-02674-5_5

de Mouzon J, Chambers GM, Zegers-Hochschild F, et al. International committee for monitoring assisted reproductive technologies world report: assisted reproductive technology 2012. Hum Reprod. 2020;35(8):1900–13. https://doi.org/10.1093/humrep/deaa090

Du K-L, Swamy MNS. *Neural Networks and Statistical Adjustments* (Issue 2007). London: Springer.

Durairaj M, Ramasamy N. A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate. Int J Control Theory Appl. 2016;9(27):255–60.

European Society of Human Reproduction and Embryology. More than 8 million babies born from IVF since the world's first in 1978: European IVF pregnancy rates now steady at around 36 percent, according to ESHRE monitoring. Science Daily. 2018:7–9.

Goldberg D, Holland J. Genetic algorithms and machine learning. Mach Leran. 1988;3:95–9.

Guh RS, Wu TC, Weng SP. Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. Expert Syst Appl. 2011;38(4):4437–49. https://doi.org/10.1016/j.eswa.2010.09.112

Hafiz P, Nematollahi M, Boostani R, Jahromi BN. Predicting implantation outcome of in vitro fertilization and intracytoplasmic sperm injection using data mining techniques. Int J Fertil Steril. 2017;11(3):184–90. https://doi.org/10.22074/ifs.2017.4882

Hassan MR, Al-Insaif S, Hossain MI, Kamruzzaman J. A machine learning approach for prediction of pregnancy outcome following IVF treatment. Neural Comput Appl. 2020;32(7):2283–97. https://doi.org/10.1007/s00521-018-3693-9

Hughes EG, King C, Wood EC. A prospective study of prognostic factors in in vitro fertilization and embryo transfer. Fertil Steril. 1989;51(5):838–44. https://doi.org/10.1016/S0015-0282(16)60676-3

Kaufmann SJ, Eastaugh JL, Snowden S, Smye SW, Sharma V. The application of neural networks in predicting the outcome of in-vitro fertilization. Hum Reprod. 1997;12(7):1454–7. https://doi.org/10.1093/humrep/12.7.1454

Khan A, Gould S, Salzmann M. Deep convolutional neural networks for human embryonic cell counting. European Conference on Computer Vision. 2016;1:651–67. https://doi.org/10.1007/978-3-319-46604-0

Kotsiantis SB, Kanellopoulos D. Data preprocessing for supervised leaning. Int J Comput Sci. 2006;1(2):1–7. https://doi.org/10.1080/02331931003692557

Louis CM, Erwin A, Handayani N, Polim AA, Boediono A, Sini I. Review of computer vision application in in vitro fertilization: the application of deep learning-based computer vision technology in the world of IVF. J Assist Reprod Genet. 2021. https://doi.org/10.1007/s10815-021-02123-2

Nanni L, Lumini A, Manna C. A data mining approach for predicting the pregnancy rate in human assisted reproduction. Comput Intell. 2010;326:97–111. https://doi.org/10.1007/978-3-642-16095-0_6

Passmore L, Goodside J, Hamel L, Gonzales L, Silberstein T, Trimarchi J. Assessing decision tree models for clinical in-vitro fertilization data. December 2003:1–15. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.7597&amp;rep=rep1&amp;type=pdf

Raef B, Ferdousi R. A review of machine learning approaches in assisted reproductive technologies. Acta Inform Med. 2019;27(3):205–11. https://doi.org/10.5455/aim.2019.27.205-211

Ratna MB, Bhattacharya S, Abdulrahim B, McLernon DJ. A systematic review of the quality of clinical prediction models in in vitro fertilisation. Hum Reprod. 2020;35(1):100–16. https://doi.org/10.1093/humrep/dez258

Simopoulou M, Sfakianoudis K, Maziotis E, et al. Are computational applications the "crystal ball" in the IVF laboratory? The evolution from mathematics to artificial intelligence. J Assist Reprod Genet. 2018;35(9):1545–57. https://doi.org/10.1007/s10815-018-1266-6

Simopoulou M, Sfakianoudis K, Antoniou N, et al. Making IVF more effective through the evolution of prediction models: is prognosis the missing piece of the puzzle? Syst Biol Reprod Med. 2018;64(5):305–23. https://doi.org/10.1080/19396368.2018.1504347

Steinberg D. CART: classification and regression trees. In: Wu X and Kumar V, eds, *The Top Ten Algorithms in Data Mining*. Milton Park: Taylor & Francis Group, LLC;2009:179–201.

Uyar A, Bener A, Ciray HN. Predictive modeling of implantation outcome in an in vitro fertilization setting. Med Decis Making. 2015;35(6):714–25. https://doi.org/10.1177/0272989X14535984

Vaegter KK, Lakic TG, Olovsson M, Berglund L, Brodin T, Holte J. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers. Fertil Steril. 2017; 107(3):641–8.e2. https://doi.org/10.1016/j.fertnstert.2016.12.005

van Loendersloot LL, van Wely M, Limpens J, Bossuyt PMM, Repping S, van der Veen F. Predictive factors in in vitro fertilization (IVF): a systematic review and meta-analysis. Hum Reprod Update. 2010;16(6):577–89. https://doi.org/10.1093/humupd/dmq015

Vogiatzi P, Pouliakis A, Siristatidis C. An artificial neural network for the prediction of assisted reproduction outcome. J Assist Reprod Genet. 2019;36(7):1441–8. https://doi.org/10.1007/s10815-019-01498-7

Zegers-Hochschild F, Mansour R, Ishihara O, et al. International committee for monitoring assisted reproductive technology: world report on assisted reproductive technology, 2005. Fertil Steril. 2014;101(2):366–78.e14. https://doi.org/10.1016/j.fertnstert.2013.10.005

Zhang S. Nearest neighbor selection for iteratively kNN imputation. J Syst Soft. 2012;85(11):2541–52. https://doi.org/10.1016/j.jss.2012.05.073