Machine learning vs. classic statistics for the prediction of IVF outcomes

Zohar Barnett-Itzhaki^{1,2,3,4} · Miriam Elbaz⁴ · Rachely Buttermann⁴ · Devora Amar⁴ · Moshe Amitay⁴ · Catherine Racowsky⁵ · Raoul Orvieto^{6,7} · Russ Hauser⁸ · Andrea A. Baccarelli⁹ · Ronit Machtinger^{6,7}

Received: 8 April 2020 / Accepted: 30 July 2020 / Published online: 11 August 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Purpose To assess whether machine learning methods provide advantage over classic statistical modeling for the prediction of IVF outcomes.

Methods The study population consisted of 136 women undergoing a fresh IVF cycle from January 2014 to August 2016 at a tertiary, university-affiliated medical center. We tested the ability of two machine learning algorithms, support vector machine (SVM) and artificial neural network (NN), vs. classic statistics (logistic regression) to predict IVF outcomes (number of oocytes retrieved, mature oocytes, top-quality embryos, positive beta-hCG, clinical pregnancies, and live births) based on age and BMI, with or without clinical data.

Results Machine learning algorithms (SVM and NN) based on age, BMI, and clinical features yielded better performances in predicting number of oocytes retrieved, mature oocytes, fertilized oocytes, top-quality embryos, positive beta-hCG, clinical pregnancies, and live births, compared with logistic regression models. While accuracies were 0.69 to 0.9 and 0.45 to 0.77 for NN and SVM, respectively, they were 0.34 to 0.74 using logistic regression models.

Conclusions Our findings suggest that machine learning algorithms based on age, BMI, and clinical data have an advantage over logistic regression for the prediction of IVF outcomes and therefore can assist fertility specialists' counselling and their patients in adjusting the appropriate treatment strategy.

Keywords IVF · Oocytes · Implantation · Machine learning · Prediction models

Introduction

Machine learning is an application of artificial intelligence (AI) that provides the ability to interpret and understand data,

and to develop prediction models based on experience. Machine learning approaches are increasingly used to construct prediction models for clinical decision-making, mostly in complex multi-variable systems [1–4].

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10815-020-01908-1>) contains supplementary material, which is available to authorized users.

✉ Zohar Barnett-Itzhaki
zoharba@ruppin.ac.il; zohari@gmail.com

¹ Public Health Services, Ministry of Health, 39 Yirmiyahu Street, 9446724 Jerusalem, Israel

² School of Engineering, Ruppin Academic Center, Emek Hefer, Israel

³ Research Center for Health Informatics, Ruppin Academic Center, Emek Hefer, Israel

⁴ Bioinformatics Department, School of Life and Health Sciences, Jerusalem College of Technology, Jerusalem, Israel

⁵ Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

⁶ Department of Obstetrics and Gynecology, Sheba Medical Center, 52561 Ramat Gan, Israel

⁷ Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

⁸ Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

⁹ Laboratory of Precision Environmental Biosciences, Department of Environmental Health Sciences, Columbia Mailman School of Public Health, New York, NY 10032, USA

Machine learning algorithms have been used recently to predict IVF outcomes based on multiple clinical variables, including ovarian stimulation variables and medical history of diseases [5–9]. Other machine learning and artificial intelligence approaches used visual images of the embryos in order to select the highest quality embryos which will hopefully lead to a viable pregnancy [10, 11].

As opposed to classic statistics, machine learning methods can model complex systems and take into account links and associations among variables. This approach may provide unbiased robust models for future predictions and enable synergism between different parameters that might indirectly be associated with the outcomes.

Prompted by the aforementioned information, we sought to examine whether machine learning methods provide an advantage over classic statistical modeling for the prediction of IVF outcomes (both intermediate and clinical) in a limited cohort of women. For this purpose, we compared the performances of two machine learning algorithms (support vector machine (SVM) with different kernel functions and artificial neural network (NN)) with a standard statistical method (logistic regression) for the prediction of IVF outcomes based on the patient age, BMI, and ovarian stimulation variables.

Methods

Institutional review board approval

The study was approved by our hospital's IRB (SMC 8707/11) and all patients signed informed consents. Most of the study population overlapped with our previous study [12].

Study population

One hundred thirty-six women undergoing fresh IVF cycles from January 2014 through August 2016, at Sheba Medical Center, Israel, a tertiary, university-affiliated hospital, participated in this prospective study. We recruited participants undergoing their 1st–5th IVF cycle and were followed through one complete fresh IVF treatment cycle. To avoid potential confounding by infertility diagnosis or treatment protocols, only women undergoing treatment due to male factor or unexplained infertility, who were oocyte donors, or couples undergoing preimplantation genetic diagnosis (PGD) for autosomal recessive diseases were recruited. All patients enrolled were treated with GnRH antagonist protocol. We excluded women > 38 years old, and those suffering from endometriosis, polycystic ovary syndrome, or decreased ovarian reserve, which might affect oocyte quality. Baseline day 3 hormonal profiles to identify normal ovarian reserve included FSH levels < 10 IU and estradiol (E2) < 200 pmol/L. In addition,

we excluded cases of social oocyte cryopreservation and also frozen IVF cycles.

Patients' characteristics and IVF outcomes

We compared the ability of the models to predict intermediate IVF outcomes (number of oocytes retrieved, number of mature oocytes, number of fertilized oocytes, number of top-quality day 3 embryos), as well as clinical IVF outcomes (positive beta-hCG, clinical pregnancies, and live birth rates).

Height and weight, measured at the start of the IVF cycle, were used to calculate BMI (kg/m^2). The number of previous pregnancies and deliveries, duration of infertility, and number of previous IVF attempts were obtained from patients' medical electronic charts.

Ovarian stimulation protocols, oocyte maturity assessment, and fertilization determination were as previously described [13]. According to our program policy, patients were scheduled for day 3 transfer. In five cases, for which day 3 occurred on a weekend, embryos were transferred on day 2. Embryos were assessed for cell number, symmetry, and fragmentation. Top-quality embryos were classified as those with 7–8 cells on day 3 and < 10% fragmentation. Slow day 3 embryos were defined as embryos with less than seven blastomeres and accelerated embryos as those who had more than eight blastomeres on day 3. In case of day 2 transfer, top-quality embryos were defined as having four equal blastomeres with less than 10% fragmentation. Positive β -hCG (i.e., successful implantation) was defined as a serum β -hCG level > 25 mIU/mL, typically measured 14–17 days after oocyte retrieval. Clinical pregnancy was defined as the presence of an intrauterine gestational sac and fetal heartbeat confirmed by ultrasound by 7 weeks of gestation. Live birth was defined as the delivery of a live neonate on or after 24 weeks of gestation.

For the models predicting clinical outcomes, i.e., IVF-positive beta-hCG, clinical pregnancy, and live birth, we excluded (1) cases of PGD to exclude any potential confounding caused by the biopsy; (2) the one cycle using oocyte donation as the oocytes were split into three different recipients; and (3) cases in which no embryos were transferred. All clinical information was abstracted from medical records.

Statistical analyses

Computational analyses were performed using MATLAB (MathWorks, Natick, MA, USA, version R2016a).

Models and machine learning algorithms

All models were built using MATLAB's Machine Learning Toolbox.

We compared the performances of two machine learning classification algorithms, support vector machine (SVM) with

different kernel functions and artificial neural network (NN), with a classical statistical model, logistic regression, for predicting IVF outcomes.

SVM is a classification algorithm that finds a hyperplane to differentiate samples from different classes. Special functions (called kernels) can be used for SVM classification in order to achieve non-linear separations. In this study, the Gaussian and the radial basis function (RBF) kernels were used, since they yielded the best performances. NNs are statistical models with a unique structure that simulates the learning dynamics of the neural cells in the brain [14]. NNs are based on a hierarchical node structure that receives multiple inputs, assigns each input a weight, and outputs a decision based on the summation of these weighted data. Logistic regression (LR) models reflect the relationship between a multi-class series of features and categorical output variables and are used to classify new samples to two or more categorical classifications.

Patient age and BMI were included in all models. We further added the following clinical features to the models compared: number of previous pregnancies, number of previous deliveries, baseline estradiol levels at day 3 of the ovarian stimulation, duration of ovarian stimulation, total dose of FSH ampules (IU) used during the cycle, number of oocytes retrieved, number of mature oocytes, number of fertilized oocytes, number of day 2/3 embryos, number of day 2/3 top-quality embryos, and maximal endometrial thickness on the day of triggering final follicular maturation. We adjusted pregnancy and live birth rate models to the number of transferred embryos. The specific data that were included in each model are shown in Table 1.

Of note, some of the clinical data were used both as predictors in several models and as target functions (predicted IVF outcomes) in other models (Table 1).

In order to find the optimal number of the networks' nodes, we ran the algorithm on the test set using 10 to 50 nodes (50 times for each number of nodes) and chose the number of nodes that maximized the performance. Due to the stochastic nature of the model, we ran the algorithms 50 times and calculated the average performances. In addition, for the machine learning analyses, we replaced the values of each of the features by their tertiles (1st, 2nd, or 3rd). In continuous target prediction models (for example, number of retrieved oocytes), we replaced each target value by their tertiles. This also enabled the categorization of the models and calculation of confusion matrices for the model evaluations. Receiver operating characteristic (ROC) curves were used to assess the performances of the clinical IVF outcome models (β -hCG, clinical pregnancy, and live birth rates) for the NN models [15].

In a further analysis of the NN model, we used feature selection based on the RReliefF algorithm (Robnik Šikonja and Kononenko 1997) to rank the features by their contribution

to the model. RReliefF algorithm scores every feature by comparing the sample similarity based on the feature to the actual class of these samples (1st, 2nd, or 3rd tertile for the intermediate IVF outcomes, or yes/no for the dichotomous IVF clinical outcomes), penalizing features according to which samples were similar while in fact they belong to different classes.

Machine learning model evaluation

Data were randomly split into a training set (70% of the data) and a test set (30%) for calculating the algorithms' accuracy. Models were created based on the training set only, and latter were tested on the test set, a process that provides an unbiased evaluation of the models. For each model, we calculated the following performance metrics on the test set: accuracy $\left(\frac{\text{True positive} + \text{True negative}}{\text{All population}}\right)$, error rate (1-accuracy), precision $\left(\frac{\text{True positive}}{\text{True positive} + \text{False positive}}\right)$, recall $\left(\frac{\text{True positive}}{\text{True positive} + \text{False negative}}\right)$, and F1 score $\left(2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}\right)$ —a score that measures the balance between precision and recall that is widely used to assess the algorithm performances. Accuracy, precision, recall, F1 score, and AUC are all in the range between 0 and 1; in general, higher scores represent better performances. Each model was run 20–50 times using 20–50 different partitions to training set and test set. We calculated the averages of each of the aforementioned metrics.

Results

Patients' demographics

The mean age of the women was 30.9 years (range 19 to 38 years) and their mean duration of infertility was 1.2 years. Sixty-one percent of the women were nulliparous, 66% of them had no previous IVF attempts, 19.5% underwent one previous IVF cycle, and another 6.5% underwent 2 previous IVF cycles. Twenty percent were current smokers. The mean and standard deviation (SD) of BMI in our cohort were $23.5 \pm 4.7 \text{ kg/m}^2$. Thirty-seven percent of the women underwent IVF for male factor, 46% for PGD, 15% unexplained, and 2% for other reasons (oocyte donors and mechanical factor). The mean number and SD of oocytes retrieved and mature oocytes were 9.8 ± 6.3 and 7.9 ± 7 respectively. Further descriptive statistics of the study cohort are shown in Supplementary Table 1.

The percent of cycles undergoing embryo transfer ($n = 72$) that resulted in positive beta-hCG, clinical pregnancy, and live birth were 39%, 35%, and 34%, respectively.

Table 1 Features used in the clinical prediction models included patient's age, BMI, number of previous pregnancies, number of previous deliveries, baseline estradiol level at day 3 (before ovarian stimulation), number of days of ovarian stimulation, and total FSH dose (IU). The additional parameters used for the six models are described below

Predicted IVF outcome	Number of oocytes retrieved	Number of mature oocytes	Number of fertilized oocytes	Number of day 3 embryos	Number of top-quality embryos	Number of transferred embryos	Maximal endometrial thickness
Number of mature oocytes	x						
Number of fertilized oocytes	x	x					
Number of top-quality embryos	x	x	x	x			
Positive beta-hCG	x	x	x	x	x	x	x
Clinical pregnancy	x	x	x	x	x	x	x
Live birth rates	x	x	x	x	x	x	x

Comparison of machine learning algorithms and logistic regression performance for predicting IVF outcomes

In order to check the relative contribution of each of the clinical features to the model and their potential dependency, we first created separate models (for each of the three algorithms—NN, SVM, and LR) to predict the number of retrieved oocytes (single-feature models). Each model was based on patient age, BMI, and one of the relevant features, including previous pregnancies, previous deliveries, baseline estradiol levels on cycle day 3, duration of stimulation, and number of FSH IUs used during ovarian stimulation. Prediction performances of the single-feature models were close to random (Supplementary Table 2).

We then compared the ability of LR, SVM, and NN models to predict both intermediate and clinical IVF outcomes, based on age, BMI, and all the relevant clinical features to the model. When using all these relevant clinical parameters, prediction accuracies and performances dramatically increased specifically for the NN model.

For the intermediate IVF outcomes, NN accuracies ranged between 0.69 and 0.88, while SVM accuracies ranged between 0.45 and 0.77. LR accuracies were lower and ranged between 0.34 and 0.74. NN maximal F1 scores were also better than those in SVM and LR (0.87, 0.74, and 0.74 for NN, SVM, and LR, accordingly). Models' performances are shown in Fig. 1.

For the clinical IVF outcomes, NN accuracies ranged between 0.85 and 0.9, while SVM accuracies range between 0.58 and 0.63 and LR accuracies ranged between 0.53 and 0.58. NN maximal F1 scores were also better than those in SVM and LR (0.89, 0.74, and 0.46 for NN, SVM, and LR, respectively). Data are shown in Fig. 2.

As machine learning algorithms, specifically NN, yielded better performance compared with classical statistics for both

intermediate and clinical IVF outcomes, we further focused on the NN. The detailed performances of LR and SVM models are presented in Supplementary Tables 3 and 4, respectively.

Table 2 presents the detailed NN model performances on the test sets for the intermediate IVF outcomes and Table 3 presents the NN model performances on the test sets for the clinical IVF outcomes (positive beta-hCG, clinical pregnancy, and live birth rates). BMI was the one variable that significantly contributed to the prediction models of the number of oocytes retrieved and number of mature oocytes, while the number of previous pregnancies significantly contributed to the prediction models of the number of oocytes fertilized and the positive beta-hCG, clinical pregnancy, and live birth rates.

We then created receiver operating characteristic (ROC) curves to further evaluate the clinical prediction models and to demonstrate their superiority over random models. We found that the area under the curve (AUC) for the prediction of positive beta-hCG increased from 0.63 for the model including age and BMI to 0.79 for the second model including the clinical parameters (Table 3). The AUC for the prediction of clinical pregnancy increased from 0.72 for the model including age and BMI to 0.86 for the second model including the clinical parameters (Table 3 and Fig. 3).

Discussion

The artificial neural network (NN) machine learning algorithm was superior in performance compared with the other algorithms evaluated, based on both unmodifiable and modifiable data, and yielded good to excellent performance (F1 scores of 0.69 to 0.89) to predict various IVF outcomes. These models demonstrated better results compared with classical statistics, such as logistic regression. Our data show that a consistent factor contributing to the prediction of the number of oocytes retrieved and the number of mature oocytes among

Fig. 1 Comparisons of intermediate IVF outcomes among the three models using BMI, age, and clinical data ($n = 136$). **a.** F1 score comparisons. **b.** Accuracy comparisons. LR, logistic regression; SVM, support vector machine; NN, neural network

our population was the woman's BMI, while a significant factor contributing to the number of fertilized oocytes, positive beta-hCG, clinical pregnancy, and live birth rates was the number of her previous pregnancies.

Predicting of IVF outcomes remains one of the challenges in the field of fertility treatments. A possible reason for this is the limited predictive accuracy and clinical utility of the current published models [16]. A recent systematic review published earlier this year evaluated the available models published. Most of the studies to date were retrospective cohort studies and more than 90% of them used logistic regression for the models. Different studies included infertile couples with various age range categories. The outcomes predicted

were heterogeneous and focused on different types of pregnancy outcomes as positive beta-hCG, clinical pregnancies, ongoing pregnancies, and live birth rates. Similar to our studies, most studies to date evaluated treatment success following one fresh embryo transfer attempt [16].

Artificial intelligence (AI) and machine learning (ML) are quickly gaining usefulness in predicting human fertility [17]. Previous studies suggested the utility of this approach to construct neural networks that include patient data is to assist clinicians when offering couples undergoing IVF personalized treatment options [18]. Previous studies that have assessed the effects of various clinical factors on IVF outcomes using artificial intelligence focused primarily on live

Fig. 2 Comparisons of clinical IVF outcomes among the three models using BMI, age, and clinical data ($n = 72$). **a.** F1 score comparisons. **b.** Accuracy comparisons. LR, logistic regression; SVM, support vector machine; NN, neural network

Table 2 Performances of NN models for IVF intermediate outcomes, based on clinical features ($n = 136$)

Intermediate IVF outcome	Set of features	Accuracy*	Recall**	Precision***	F1 score****
Number of retrieved oocytes	Age + BMI	52.2%	0.52	0.57	0.49
	Clinical data + age + BMI	69.4%	0.7	0.71	0.69
Number mature oocytes	Age + BMI	54.2%	0.53	0.59	0.51
	Clinical data + age + BMI	88.1%	0.88	0.88	0.87
Number of fertilized oocytes	Age + BMI	53.5%	0.52	0.56	0.5
	Clinical data + age + BMI	77.5%	0.78	0.79	0.77
Number of top-quality embryos	Age + BMI	52.9%	0.52	0.57	0.5
	Clinical data + age + BMI	85.7%	0.86	0.86	0.85

*Accuracy = $\frac{\text{True positive} + \text{True negative}}{\text{All population}}$; **Recall = $\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$; ***Precision = $\frac{\text{True positive}}{\text{True positive} + \text{False positive}}$; ****F1 score = $2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$

Clinical data included the following: number of previous pregnancies, number of previous deliveries, baseline estradiol level at day 3 (before ovarian stimulation), number of days of ovarian stimulation, and total FSH dose (IU) as well as those described in Table 1

Table 3 Performances of NN models for IVF clinical outcomes, based on clinical features ($n = 72$)

IVF outcome	Set of features	Accuracy*	Recall**	Precision***	F1 score****	AUC
a. Positive beta-hCG	Age + BMI	73.8%	0.77	0.75	0.73	0.63
	Clinical data + age + BMI+ transferred embryos	85.1%	0.86	0.86	0.84	0.79
b. Clinical pregnancy	Age + BMI	75.5%	0.75	0.78	0.74	0.72
	Clinical data + age + BMI+ transferred embryos	89.6%	0.90	0.90	0.89	0.86
c. Live birth	Age + BMI	71.8%	0.71	0.75	0.69	0.68
	Clinical data + age + BMI+ transferred embryos	87.4%	0.88	0.88	0.86	0.77

*Accuracy = $\frac{\text{True positive} + \text{True negative}}{\text{All population}}$, **Recall = $\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$, ***Precision = $\frac{\text{True positive}}{\text{True positive} + \text{False positive}}$, ****F1 score = $2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$

birth rates. Vogiatzi et al. collected 118 parameters for each of 426 IVF/ICSI cycles of women < 43 years old, treated either with GnRH antagonist or GnRH agonist protocols, and used neural networks to correlate these parameters with the outcome of live birth [9]. Age at menarche was highly correlated with live birth rates while female age and number of top-quality embryos were somewhat correlated with live birth after IVF.

As opposed to classical statistical models as logistic regression, machine learning algorithms can take into consideration complex associations between different parameters and can consequently better utilize the synergism between these associated parameters (including parameters that are not directly associated with the outcomes may be used in NN models and improve performances). This may explain their superiority over classic statistical analyses.

Computational models can identify complex, non-linear associations that cannot be detected by other approaches. Our study emphasizes also the importance of modifiable variables as contributors to IVF success. Our cohort included mainly women with normal BMI (53%) or overweight (22%). However, 10% were obese and 6% were morbid obese. BMI had a key role in the prediction of the number

of retrieved oocytes and the number of mature oocytes. Such associations with BMI are in line with a previous study, testing follicular extracellular vesicle miRNAs in the same population, in which we showed that 1-unit increase in BMI is associated with altered follicular fluid expression of extracellular vesicle-linked miRNAs. These changes might influence follicular and oocyte developmental pathways [12].

The main limitation of this study is the relatively limited sample of participating women. Our study population was carefully enrolled (relatively young patients, similar ovulation induction protocols, as well as exclusion of infertility diagnoses that might affect oocyte quality). Although these exclusion criteria resulted in a relative small sample, they enabled noise reduction and enabled building robust models with good performances. A second limitation of the current study is unavailability of external data to validate the results. However, the training-test approach in which the models are solely built based on the training set and are tested on a different set of samples enable reliable results regarding the models' performances.

This study has several strengths. First, we used a dataset collected from a homogeneous population of relatively young women undergoing the same IVF protocol and excluding

Fig. 3 ROC curves for clinical pregnancy predictions (NN model)

ovarian factors as diminished ovarian reserve, endometriosis, and polycystic ovary syndrome which might affect IVF outcomes, and we undertook assessment of detailed intermediate outcomes throughout the IVF process. Second, we were able to show reliable models with good to excellent performances using machine learning approaches.

The ability to predict IVF outcomes is one of the major challenges for both patients and clinicians in the field of fertility treatments. The models developed in this work have international relevance and can be used to assess IVF outcomes based on age, BMI, and clinical data. Hopefully, the accumulation of further data from different sources can further improve machine learning models and therefore enable personalized treatment.

Acknowledgments We would like to thank Dr. Tamar Berman for her insightful comments.

Availability of data and material Data will be shared upon request.

Funding information This work was supported by the National Institutes of Health (grant numbers P30ES00002, R21ES024236, and K99ES026648) and by the Environment and Health Fund, Israel (grant award no. RPGA1301).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Sheba (8707/11).

Consent to participate All patients signed an informed consent according to the IRB approval.

Consent for publication Yes

Code availability Not applicable

References

- Adamker G, Holzer T, Karakis I, Amitay M, Anis E, Singer SR, et al. Prediction of shigellosis outcomes in Israel using machine learning classifiers. *Epidemiol Infect.* 2018;146(11):1445–51.
- Deo RC. Machine learning in medicine. *Circulation.* 2015;132(20):1920–30.
- Jovic S, Miljkovic M, Ivanovic M, Saranovic M, Arsic M. Prostate cancer probability prediction by machine learning technique. *Cancer Invest.* 2017;35(10):647–51.
- Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med.* 2016;375(13):1216–9.
- Blank C, Wildeboer RR, DeCruo I, Tillemann K, Weyers B, de Sutter P, et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. *Fertil Steril.* 2019;111(2):318–26.
- Güvenir HA, Misirli G, Dilbaz S, Özdegirmenci O, Demir B, Dilbaz B. Estimating the chance of success in IVF treatment using a ranking algorithm. *Med Biol Eng Comput.* 2015;53(9):911–20.
- Hafiz P, Nematollahi M, Boostani R, Namavar JB. Predicting implantation outcome of in vitro fertilization and intracytoplasmic sperm injection using data mining techniques. *Int J Fertil Steril.* 2017;11(3):184–90.
- Uyar A, Bener A, Ciray HN. Predictive modeling of implantation outcome in an in vitro fertilization setting: an application of machine learning methods. *Med Decis Making.* 2015;35(6):714–25.
- 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.
- Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit Med.* 2019;2:21.
- Manna C, Nanni L, Lumini A, Pappalardo S. Artificial intelligence techniques for embryo and oocyte classification. *Reprod Biomed Online.* 2013;26(1):42–9.
- Martinez RM, Baccarelli AA, Liang L, Dioni L, Mansur A, Adir M, et al. Body mass index in relation to extracellular vesicle-linked microRNAs in human follicular fluid. *Fertil Steril.* 2019;112(2):387–96 e3.
- Machtinger R, Gaskins AJ, Racowsky C, Mansur A, Adir M, Baccarelli AA, et al. Urinary concentrations of biomarkers of phthalates and phthalate alternatives and IVF outcomes. *Environ Int.* 2018;111:23–31.
- Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet.* 1995;346(8982):1075–9.
- Robnik Šikonja M, Kononenko I. An adaptation of relief for attribute estimation in regression. In: Fisher DH (editor). *Machine Learning: Proceedings of the Fourteenth International Conference; 1997 July 8–12; Nashville, (TN): Morgan Kaufmann, 1997: 296–304.*
- 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.
- Curchoe CL, Bormann CL. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. *J Assist Reprod Genet.* 2019;36(4):591–600.
- Siristatidis C, Vogiatzi P, Pouliakis A, Trivella M, Papantoniou N, Bettocchi S. Predicting IVF outcome: a proposed web-based system using artificial intelligence. *In Vivo.* 2016;30(4):507–12.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.