REVIEW ARTICLE
https://doi.org/10.5653/cerm.2023.06009
pISSN 2233-8233 • eISSN 2233-8241
Clin Exp Reprod Med 2024;51(1):1-12

CERM

Criteria for implementing artificial intelligence systems in reproductive medicine

Enric Güell[1,2]

[1]CONSULTFIV, Valls; [2]Procrear, Reus, Spain

This review article discusses the integration of artificial intelligence (AI) in assisted reproductive technology and provides key concepts to consider when introducing AI systems into reproductive medicine practices. The article highlights the various applications of AI in reproductive medicine and discusses whether to use commercial or in-house AI systems. This review also provides criteria for implementing new AI systems in the laboratory and discusses the factors that should be considered when introducing AI in the laboratory, including the user interface, scalability, training, support, follow-up, cost, ethics, and data quality. The article emphasises the importance of ethical considerations, data quality, and continuous algorithm updates to ensure the accuracy and safety of AI systems.

Keywords: Artificial intelligence; Criteria; Deep learning; Implementation; Machine learning; Reproductive medicine

Introduction

Artificial intelligence (AI) algorithms have become ubiquitous in our lives, and the field of assisted reproductive technology (ART) is no exception. In recent years, increasingly many publications in scientific journals and conferences have highlighted the various applications of AI in reproductive medicine [1,2]. These applications span a wide range of areas within the field of reproductive medicine [3-5]. As embryologists, as well as physicians, we have the duty to keep abreast of the existing technologies, and above all, their function and results, before accepting the incorporation of any new tool in clinical practice. The present work aims to provide key concepts to be taken into consideration when considering integrating AI systems into reproductive medicine practices.

Artificial intelligence in assisted reproductive technology

Among the numerous published algorithms, we can find predictive models for embryo transfer outcomes on day 2/3 [6] and blastocyst stage [7,8], sperm selection by image recognition correlated with fertilization and blastocyst formation [9], prediction of obtaining spermatozoa from testicular biopsies [10], non-invasive oocyte scoring on two-dimensional images [11], cytoplasmic recognition of the zygote [12], morphokinetic automated annotation of the embryo [13-15], automated blastocyst quality assessment [16], embryo implantation potential via morphokinetic biomarkers [17], euploidy prediction using metabolic footprint analysis [18], ranking for embryo selection [19-25], blastocoel collapse and its relationship with degeneration and aneuploidy [26], morphokinetics and clinical features for the prediction of euploidy [27], prediction of aneuploidy or mosaicism using only patients' clinical characteristics [28], tracking of menstrual cycles and prediction of the fertile window [29], control of culture conditions and quality control of embryologist performance [25,30], intrauterine insemination success [31], computer decision support for ovarian stimulation [32], prediction for the day of triggering [33,34], and follicle-stimulating hormone dosage prediction for ovarian stimulation [35]. All the mentioned references are depicted in

Received: March 15, 2023 · Revised: August 12, 2023 · Accepted: August 31, 2023
Corresponding author: **Enric Güell**
CONSULTFIV, Pg. Estació 18 5-5, 43800 Valls, Spain
Tel: +34-637250080 Fax: +34-637250080 E-mail: eguellp@gmail.com

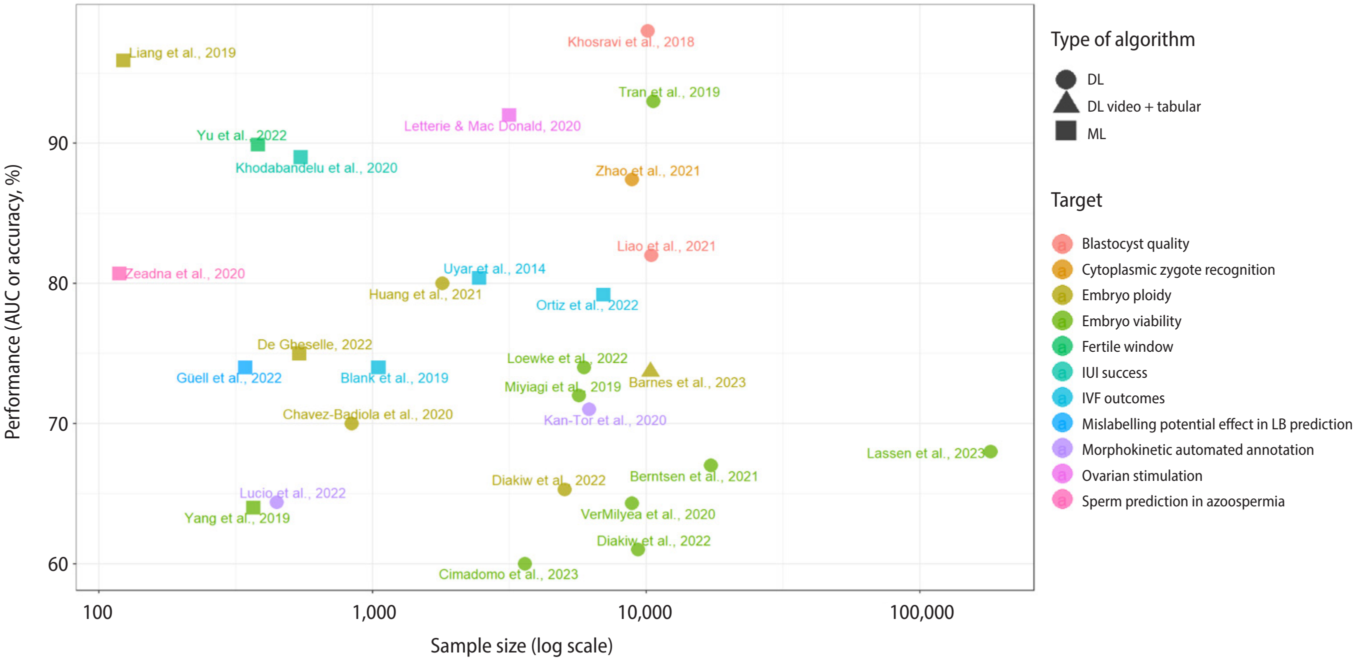

Figure 1. The main AI algorithms published ordered by sample size (log scale) and the performance metric given by the authors (area under the curve [AUC] or accuracy), grouped by type of machine learning (ML) algorithm (shape) and target (colour). The AUC values, inherently bounded within the range from 0 to 1, have been converted to percentages by multiplying them by 100 in this graph for enhanced interpretability. DL, deep learning; IUI, intrauterine insemination; IVF, *in vitro* fertilization; LB, live birth.

Figure 1. Machine learning models are listed in Table 1 [6,7,10,17,18,27-29,31-37], while those corresponding to the deep learning subset can be found in Table 2 [8,9,11-16,19-24,26,38-43]. In these tables, the AI models are described with their sample size, results and limitations. The main limitation of all studies was their retrospective nature. A limited sample size, imbalanced dataset, and lack of multicenter evaluation were also common limitations found in the literature review.

Commercial platforms or in-house algorithms

The AI systems used in *in vitro* fertilization (IVF) clinics can be categorised into two types: commercial products and self-developed in-house solutions. While cloud-based systems can offer advantages for IVF clinics with lower workloads, such as leveraging data from other clinics, they may face challenges in maintaining predictive accuracy due to interference from individual clinic protocols or conditions. Notable examples of cloud-based products include Embryo Ranking Intelligent Classification Algorithm (ERICA) [19], intelligent Data Analysis Score (iDAScore) [23], and Life Whisperer [20].

In contrast, adopting an in-house approach could offer certain advantages, such as greater control and customisation over the AI system and its workflow as well as the possibility to test own ideas without having to wait for commercial releases. Single-center studies such as Zeadna et al. [10] or De Gheselle et al. [27] represent this approach to AI in IVF.

Requirements for implementing new AI systems in the laboratory

Prior to introducing a new AI system—or any other technique—it is essential to ensure that it satisfies certain criteria in a laboratory setting. At least one of the following criteria should be met for the new technique to be considered suitable: the candidate AI system should have the ability to improve results, such as the live birth (LB) rate, time to pregnancy, or any other key performance indicator. If the results are not worsened, other criteria to be met could include making work easier and more efficient, saving time and resources, offering greater safety through an improved error detection, or providing better explainability.

Factors to consider when introducing AI in the laboratory

There are several factors that cannot be overlooked when considering the integration of a new system into the laboratory. These fac-

Table 1. Artificial intelligence models for assisted reproduction techniques

Study	Target	No. of features input	Dataset	Type of input	Results	Limitations
Uyar et al. (2015) [6]	IVF outcomes (implantation)	11	2,453	Patient metadata and embryo morphological characteristics	Accuracy = 80.4%; Sensitivity = 63.7%	Retrospective study, embryo transfers performed at cleavage stage (D+2/3), manual assessment, imbalanced dataset, LB not used as an endpoint, lack of multi-center evaluation.
Blank et al. (2019) [7]	IVF outcomes (implantation)	32	1,052	Patient metadata	AUC = 0.74; Sensitivity = 0.84; Specificity = 0.58	Retrospective study, limited sample size, lack of external validation, balanced training set but imbalanced testing set, LB not used as an endpoint.
Goyal et al. (2020) [37]	IVF outcomes (LB)	25	141,160	Patient metadata	AUC = 0.846; Recall = 76%; Precision = 77%	Retrospective study, imbalanced dataset adjusted by downsampling, limited generalizability due to specific population data, limited factors in the dataset.
Ortiz et al. (2022) [28]	IVF cycle outcomes (mosaicism/aneuploidy)	29	6,989	Patient metadata classified into six groups: general, maternal, paternal, couple-related, IVF cycle-related, and embryo-related	AUC_aneuploidy = 0.792; AUC_mosaicism = 0.776	Retrospective study, biased dataset as only PGT-A embryos included, imbalanced dataset, lack of external validation.
Zeadna et al. (2020) [10]	Sperm prediction in azoospermia	14	119	Patient metadata (hormonal levels, age, body mass index, histopathology, varicocele, etc.)	AUC = 0.807	Retrospective study, limited generalizability due to population heterogeneity of non-obstructive azoospermia patients, limited sample size, only TESE used, imbalanced dataset, lack of multi-center evaluation.
Guell Penas et al. (2022) [36]	Mislabelling potential effect in LB prediction	4	343	Morphokinetic parameters	AUC_aneuploid = 0.74; AUC_KIDn = 0.59	Retrospective study, PGT-A on D+3, no mosaicism considered, limited sample size, lack of multi-center evaluation, manually kinetic parameters.
Yang et al. (2019) [17]	Embryo viability (implantation, LB)	5	367	Morphokinetic parameters	AUC_implantation = 0.69; AUC_live birth = 0.64	Conventional IVF and ICSI embryo morphokinetic parameters merged without adjusting t0 for conventional IVF sperm lagging or using intervals, imbalanced dataset, limited sample size, lack of multi-center evaluation, potential mislabelled embryos in non-implanted group, retrospective study.
Liang et al. (2019) [18]	Embryo ploidy status	1	123 embryos (1,107 Raman spectra)	Raman spectra	Accuracy = 95.9%	Retrospective study, limited sample size, LB not used as an endpoint, lack of multi-center evaluation.
De Gheselle et al. (2022) [27]	Embryo ploidy status	85	539	Morphokinetic parameters, standard development features and patient metadata	AUC = 0.75; Accuracy = 71%	Retrospective study, limited sample size, LB not used as an endpoint, lack of multi-center evaluation.
Yu et al. (2022) [29]	Fertile window	2	382	Basal body temperature and heart rate	AUC = 0.899; Accuracy = 87.5%, Sensitivity = 69.3%; Specificity = 92%	Retrospective study, lower predictions for irregular menstruators (AUC = 0.725), lack of multi-center evaluation.
Khodabandelu et al. (2022) [31]	IUI success	8	546	Patient metadata, sperm sample data, antral follicle counting	Gmean, AUC, Brier values of 0.80, 0.89, and 0.129, respectively	Retrospective study, imbalanced dataset, lack of multi-center evaluation.

(Continued to the next page)

Table 1. Continued

Study	Target	No. of features input	Dataset	Type of input	Results	Limitations
Letterie et al. (2020) [32]	Ovarian stimulation day-to-day decision-making tool	4	3,159	Stimulation control data (oestradiol level, follicle measurement, cycle day and recombinant FSH dose)	Accuracy_continue = 0.92; Accuracy_trigger = 0.96; Accuracy_dosage = 0.82; Accuracy_days = 0.87	Retrospective study, lack of multi-center evaluation, manual assessment of ultrasound observations, imbalanced dataset.
Hariton et al. (2021) [33]	Ovarian stimulation trigger decision-making tool	12	7,866	Number of follicles 16–20 mm in diameter, the number of follicles 11–15 mm in diameter, and oestradiol level, age, body mass Index, protocol type	Average outcome improvement in total 2PNs and usable blastocysts compared with the physician's decision	Retrospective study, long period (> 10 years).
Fanton et al. (2022) [34]	Gonadotrophin starting dose	4	18,591	Age, body mass index, anti-Müllerian hormone, antral follicle count	Mean absolute error of 3.79 MII; r^2 for MII prediction = 0.45	Retrospective study, model based on U.S. population, dismissed confounding factors such as dose adjustments and timing of trigger, exclusion of cycles with missing data.
Correa et al. (2022) [35]	Ovarian stimulation first FSH dosage	5	3,487	Age, body mass index, anti-Müllerian hormone, antral follicle count, previous LB	Model's score approaches best possible dose more times than clinicians.	Retrospective study, tendency of the model to overdose some patients due to underrepresentation of hyper-responder, limited generalizability due to specific population data, lack of multi-center evaluation.

The main limitations, results, and sample size are presented in this table.
IVF, *in vitro* fertilization; LB, live birth; AUC, area under the curve; PGT-A, preimplantational genetic testing for aneuploidy; TESE, testicular sperm extraction; KIDn, negative known implantation embryos (non-implanted embryos); ICSI, intracytoplasmic sperm injection; FSH, follicle-stimulating hormone; IUI, intrauterine insemination; PN, pronucleate; MII, metaphase II oocyte.

tors must be carefully evaluated before making a decision. When introducing an AI system, the following factors must be taken into account:

1. User interface

The user interface (the visual display on the screen) should be easy to understand and navigate.

2. Scalability

The system should be capable of adapting to the laboratory's needs, including the volume of data and users, as well as being integrated into the laboratory's workflow and protocols. If the AI platform cannot be adapted to the laboratory's existing workflow, it is necessary to evaluate the impact of adapting the lab workflow and the potential benefit of using that AI platform.

3. Training

The manufacturer should offer information regarding the required training for users and how it will be delivered.

4. Support

The manufacturer should specify the type of technical support offered, who will be responsible in case of failure, and what the response time will be.

5. Follow-up

As AI systems continuously learn, it is crucial to ensure that the algorithms are updated to accommodate new data. The manufacturer should provide information about the maintenance and monitoring plan to ensure that the system continues to provide accurate and unbiased results.

6. Cost

The cost of a system should be considered in relation to the center's budget and investment capacity.

7. Ethics

To ensure that an AI system is ethically sound, it is important to evaluate its impact on patient care and outcomes. The system should not only improve patient outcomes but also avoid any harm or neg-

Table 2. Deep learning models for assisted reproduction techniques

Study	Target	Dataset	Type of input	Results	Limitations
Mendizabal-Ruiz et al. (2022) [9]	Sperm selection (fertilization and blastocyst formation)	383	Video	Software's scores related with fertilization ($p=0.004$) and blastocyst formation ($p=0.013$)	Retrospective study, limited sample size, lack of standardized protocols for imaging, lack of multi-center evaluation.
Nayot et al. (2021) [11]	Oocyte quality and blastocyst development	16,373	Single image	SRs related with blastocyst rate: (0–2.5) = 16%; (2.6–5) = 36.9%; (5.1–7.5) = 44.2%; (7.6–10) = 53.4%	Retrospective study, male factor was not taken into consideration, lack of standardized protocols for imaging, limited generalizability due to specific population data, LB not used as an endpoint.
Zhao et al. (2021) [12]	Cytoplasmic zygote recognition	8,877	Single image	AUC = 0.874	Retrospective study, limited sample size, requirement of high-quality images and standardization protocols for imaging, lack of external validation.
Kan-Tor et al. (2020) [13]	Morphokinetic automated annotation (blastocyst, implantation)	6,200 blastocyst, 5,500 implantation	Video	AUC_blastocyst = 0.83; AUC_implantation = 0.71	Retrospective study, not enough information about imbalancing assessment, LB not used as an endpoint, potential mislabelled embryos in non-implanted group.
Feyeux et al. (2020) [14]	Morphokinetic automated annotation	701	Video	Manual vs. automated annotation concordance, $r^2=0.92$	Retrospective study, lack of multi-center evaluation, only one focal plane.
Lucio et al. (2022) [15]	Morphokinetic automated annotation (ploidy, implantation)	448	Video	Concordance correlation coefficient ranging from tPNf = 0.813 to tSB = 0.947; AUC_plastocyst = 0.814; AUC_ploidy = 0.644	Retrospective study, imbalanced dataset, LB not used as an endpoint, mosaic score not significantly predictive, potential mislabelled embryos in non-implanted group.
Khosravi et al. (2018) [16]	Blastocyst quality	10,148	Single image	AUC = 0.98; accuracy = 96.94%; IR Good-Morph and < 37 years = 66.3%; IR Poor-Morph and ≥ 41 years = 13.8%	Retrospective study, lack of external validation, possibly limited sample size, LB not used as an endpoint.
Liao et al. (2021) [8]	Blastocyst stage and quality	10,432	Video	AUC = 0.82; accuracy = 78.2%	Retrospective study, lack of multi-center evaluation, clinical characteristics not taken into account, only one focal plane of 3D embryos, LB not used as an endpoint.
VerMilyea et al. (2020) [42]	Embryo viability (blastocyst, implantation)	8,886	Single image	Accuracy = 64.3%; sensitivity = 70.1%; specificity = 60.5%; AI improvement vs. embryologists' accuracy = 24.7%	Retrospective study, model only trained on day 5 transferred embryos, LB not used as an endpoint, funded by commercial companies, potential mislabelled embryos in non-implanted group.
Tran et al. (2019) [41]	Embryo viability (implantation)	10,638	Video	AUC = 0.93	Retrospective study, imbalanced dataset, arrested embryos included, random prediction for not arrested embryos, limited sample size, LB not used as an endpoint, funded by commercial companies, potential mislabelled embryos in non-implanted group.
Diakiw et al. (2022) [21]	Embryo viability (implantation)	9,359	Single image	AUC = 0.61; accuracy = 61.8%	Retrospective study, use of simulated cohort ranking analyses, LB not used as an endpoint.
Berntsen et al. (2022) [23]	Embryo viability (implantation)	17,249	Video	AUC_implantation = 0.67; AUC_all = 0.95	Retrospective study, imbalanced dataset, LB not used as an endpoint, funded by commercial companies, potential mislabelled embryos in non-implanted group.
Loewke et al. (2022) [40]	Embryo viability (implantation)	5,923	Single image	AUC = 0.74. Score difference of > 0.1 related to higher pregnancy rates.	Retrospective study, limited sample size, lack of standardized protocols for imaging.

(Continued to the next page)

Table 2. Continued

Study	Target	Dataset	Type of input	Results	Limitations
Theilgaard Lassen et al. (2023) [24]	Embryo viability (implantation)	181,428	Video	AUC_blind = 0.68; AUC_D+5 = 0.707; AUC_D+3 = 0.621; AUC_D+2 = 0.669; AUC_all = 0.954	Retrospective study, imbalanced dataset, upsampling may increase sample bias, AUC_all included arrested embryos, funded by commercial companies, potential mislabelled embryos in non-implanted group.
Miyagi et al. (2019) [38]	Embryo viability (LB)	5,691	Single image	Comparison AI vs. conventional embryo evaluation: 0.64/0.61, 0.71/0.70, 0.78/0.77, 0.81/0.83, 0.88/0.94, and 0.72/0.74 for the age categories < 35, 35–37, 38–39, 40–41, and ≥ 42 years and all ages, respectively.	Retrospective study, imbalanced dataset, limited sample size, lack of standardized protocols for imaging, lack of multi-center evaluation, potential mislabelled embryos in non-implanted group.
Cimadomo et al. (2023) [43]	Embryo viability (implantation, LB)	3,604	Video	AUC_euploid = 0.6; AUC_LB = 0.66	Retrospective study, imbalanced dataset, potential mislabelled embryos in non-implanted group.
Diakiw et al. (2022) [20]	Embryo ploidy status	5,050	Video	Accuracy = 65.3%; sensitivity = 74.6%; Accuracy_cleansed[a)] = 77.4%	Retrospective study, LB not used as an endpoint, lack of standardized protocols for imaging, possibly limited sample size.
Huang et al. (2021) [39]	Embryo ploidy status	1,803	Patient metadata and video	AUC = 0.8	Retrospective study, limited sample size, imbalanced dataset, manually kinetic parameters, lack of multi-center evaluation, model based only on PGT-A patients, LB not used as an endpoint.
Chavez-Badiola et al. (2020) [19]	Embryo ploidy status and viability (implantation)	840	Single image	Accuracy = 70%; PPV = 79%, NPV = 66%. Higher ranking metric (NDCGs) than random selection	Retrospective study, imbalanced dataset, lack of multi-center evaluation, limited sample size, LB not used as an endpoint, potential mislabelled embryos in non-implanted group.
Barnes et al. (2023) [22]	Embryo ploidy status	10,378	Patient metadata, video, morphokinetics, embryo grading	AUC = 0.737, accuracy = 65.7%, aneuploid predictive value = 82.3%	Retrospective study, biased dataset as only PGT-A embryos included, lack of standardized protocols for imaging, manually annotated morphokinetics and morphological assessments, differences in mosaic reporting across different genetic laboratories, mosaicism was not considered during model development, LB not used as an endpoint.
Cimadomo et al. (2022) [26]	Blastocoel collapse and its relationship with degeneration and aneuploidy	2,348	Video	Degeneration and aneuploidy rates directly related to number of collapses.	Retrospective study, limited sample size, mosaicism could not be reliably assessed, no differences in LBR.

The main limitations, results, and sample size are presented in this table.

SR, score range; LB, live birth; AUC, area under the curve; tPNf, time of pronuclear fading; tSB, time of starting blastulation; IR, implantation rate; 3D, three-dimensional; AI, artificial intelligence; PGT-A, preimplantational genetic testing for aneuploidy; PPV, positive predictive value; NPV, negative predictive value; NDCG, normalised discontinued cumulative gain.

[a)] Cleansed, refers to a data pre-processing technique aiming to reduce noisy labels.

ative impact on the patient. Moreover, the manufacturer should have measures in place to ensure the confidentiality and security of patient data, such as the ISO 27002-2021 and IEC 62304 standards. The most important ethical issue is the lack of randomised controlled trials. It is premature to implement a technology in the clinical setting before the trial results are made available [44]. The nature of the mathematical algorithms performed during the AI process leads to a spectrum of transparency, ranging from the most interpretable models, such as linear regression-based algorithms, to the most cryptic models, also called black-box, such as neural networks. It is important to know the risks, side effects, benefits and the confidence of each clinical decision before delegating the decision-making process to machines. While transparent models enhance clinical decision-making, black-box systems replace human decisions, leading to uncertainty about the responsibility for treatment success. Black-box algorithms could build predictive models biased by cofounders, and the error-checking processes of each prediction could go unnoticed by human operators [44]. Moreover, opaque models could increase the risk of imbalanced outcomes. For instance, if there exists a correlation between embryo quality assessed by AI and gender, there could be an intrinsic imbalance that could take more time to detect than in interpretable models.

8. Data quality

The quality of data refers to the data's accuracy, completeness, timeliness, relevance, consistency, and reliability. It is crucial for an AI system to have access to high-quality data to provide accurate and reliable results. If the data used for building the model are not reliable and generalisable, then the AI model will fail when applied to new data in the near future. Some models are based on a concrete and certain population, and if data across populations are not as homogeneous, then the model will not be accurate enough. Furthermore, in embryology, confounding factors such as age should not be used as predictors in embryo quality models if it is desired to develop an embryo quality model instead of an age-based predictive model [44], as the AI algorithm could base its prediction mostly on data included in the age variable with no importance for embryonic features.

9. Performance

Performance refers to the effectiveness and efficiency of an AI system in achieving its intended objectives, such as accuracy, speed, and reliability. The system's performance should be evaluated based on relevant metrics and benchmarks to ensure that it meets the desired standards.

Data annotation

The source of data is crucial in data annotation. The origin of data can vary (tabular, images, videos, audio, the outcome of a previous AI algorithm, etc.), and the annotation of data is expected to be more effective when automated, since automation removes the subjectivity of human-annotated parameters. However, the effectiveness of automated versus manual annotation depends on the degree of intra-individual and inter-individual variability for the target variable when annotated by humans and the reliability of the automatic annotation methods [45,46]. Features with higher variability or lower reliability can lead to lower performance of predictive models, since AI may use different values for data that are actually equivalent. Including such features in the models can introduce noise or inconsistencies, affecting the accuracy of predictions and the model's overall performance. Determining whether manual or automated annotation is more suitable depends on each specific case. Factors such as data complexity, available resources, and the desired level of accuracy need to be considered. Manual annotation can provide more accurate and reliable results, but can be time-consuming and introduce human biases. Automated annotation methods can be more efficient and scalable, but may be less accurate or reliable, especially in cases with noisy data or lack of proper validation.

It is not always possible for all values in a database to be filled. Not available (NA) values represent a problem when building AI algorithms and require proper handling. Several options exist for managing missing values. Some common approaches include discarding observations with NA values, imputing missing values using methods such as mean or median imputation, or utilising other AI algorithms such as k-nearest neighbour for imputation, as well as directly excluding the feature with NA values.

Machine learning techniques are also sensitive to data points that deviate significantly from the majority of the data (outliers). Managing outliers involves deciding whether to integrate them into the analysis or discard them.

Therefore, careful consideration is required when dealing with NA values and outliers. The choice of appropriate strategies for managing them depends on the specific context, the nature of the data, and the objectives of the analysis.

Risk factors affecting data quality in model design

Each predictive model has its unique characteristics and objectives, and is based on a specific experimental design that includes certain factors as inclusion and exclusion criteria. It is crucial to carefully review the experimental design, as there could be potential risks that may affect the quality of data used in the model. One such risk would be the possibility of data bias due to the inclusion criteria, which could compromise the generalisability of the results, particularly if there were confounding factors affecting predictive variables [47,48]. Three additional pitfalls to consider, as described by Curchoe et al. [49], are small sample sizes, imbalanced datasets, and limited performance metrics.

Furthermore, in classification cases, there could exist a risk of mislabelling in the output variable. Mislabelling occurs when the categorical variable has incorrect labels for some of the data points. It is important to be aware of this risk, as the inclusion of mislabelled data decreases accuracy [50,51]. A potential example of mislabelling in embryology is evident in two embryo selection models with different labels for classification. One model compares implanted or LB embryos versus non-implanted or non-live birth (NLB) embryos [38], while the other compares euploid versus aneuploid embryos [39]. In the LB versus NLB comparison, it is important to carefully consider the potential for mislabelling, as high-potential embryos with a negative outcome due to reasons unrelated to the embryo could be incorrectly labelled as NLB, which may negatively impact the performance of machine learning and deep learning algorithms [36,40,52]. Additionally, in ploidy models, undetected mosaicism [53] can also lead to mislabelling. Moreover, the "Schrödinger embryo" paradox makes it impossible to assess the genetic status of the inner cell mass and trophectoderm until the whole embryo has been donated for research. Once an embryo has been donated, it becomes impossible for it to achieve LB, and its real potential for viability will remain unrealised. Besides, the algorithms' performance may be distorted depending on the inclusion criteria in each experimental design. There is a risk of including embryos with low viability potential, those that have not yet been transferred, or even euploid embryos that were not cryopreserved due to low quality [54]. Specifically, Tran et al. [41] reported that the area under the curve (AUC) could be inflated by including many arrested embryos in the sample used to compute it. That predictive model could be considered proper for justifying automation for the quality assessment of arrested embryos, although random choice was supposed to be used for non-arrested embryos [55].

Machine and deep learning modelling

Machine and deep learning modelling refer to the process of creating and training mathematical models that can automatically identify patterns and make predictions or decisions based on data. Deep learning is included in the broader category of machine learning category. These models are built using algorithms and statistical techniques that allow computers to learn from large datasets and improve their performance over time [56]. To emphasise the main differences, it is worth noting that machine learning typically requires fewer data points and provides greater interpretability than deep learning. As a rule of thumb, the sample size should be at least 10 times the number of parameters in an algorithm, and it is generally easier to determine this value for machine learning models than for deep learning models [17].

There are two primary types of machine learning algorithms: supervised and unsupervised. On the one hand, supervised learning is an approach in which a model is trained using labelled data. After introducing input features (independent variables) along with corresponding target labels (dependent variable), supervised learning tries to learn a function or a relationship between the input features and the target labels. Once trained, the model can make predictions or classify new instances based on the input features. Supervised learning is commonly used in prediction and classification problems, where the objective is to predict a specific outcome or category, although numerical values can also be predicted through regression models. Decision trees, scoring systems, generalised additive models, and case-based reasoning are among the primary techniques used in various supervised learning algorithms [57]. Each algorithm has its own specific characteristics and uses. Linear regression involves fitting a linear equation to the data, enabling the prediction of continuous target variables [35]. Logistic regression is mainly used for binary classification tasks, although it could also be useful for multi-class problems, by modelling the probability of an event occurring based on input features [31]. Recursive partitioning is a technique commonly used in decision trees, where the data are recursively split into subsets based on certain conditions of features [57]. Random forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting [17,31]. The k-nearest neighbour method classifies or predicts the value of a data point based on the values of its k-nearest neighbours in the feature space [34,57]. Gradient boosting is an ensemble technique that builds a strong predictive model by iteratively combining multiple weak models, often decision trees, to correct errors made by the previous models [31]. Support vector machines construct hyperplanes in a high-dimensional feature space to separate different classes or estimate continuous target variables [31,57]. Neural networks are complex and versatile machine learning algorithms capable of handling various tasks, including classification, regression and pattern recognition. They are inspired by the structure of the human brain. Image recognition models are based on this type of algorithms [13,16,19,20].

On the other hand, unsupervised learning is employed in situations where the training data lack pre-existing labels or outcomes. Its objective is to discover patterns or structures inherent in the data without explicit guidance and to uncover similar groups or clusters of data. This type of learning is useful for exploring and comprehending the underlying structure in data and identifying hidden patterns. Clustering algorithms and dimensionality reduction methods are widely used in the field of unsupervised learning. K-means is a popular clustering algorithm aiming to divide a dataset into distinct groups or clusters based on similarity. The algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroids until convergence [17]. Principal component analysis (PCA) is a dimensionality reduction technique that transforms a high-dimensional dataset into a lower-dimensional space by identifying the principal components that capture the most significant variance in the data. These principal components are orthogonal and ordered in terms of their explanatory power. PCA is useful for simplifying complex datasets, visualising data in lower dimensions, and identifying the most important features driving variability in the data [56].

Thus, the algorithms used in assisted reproduction to predict categories using labelled data are of the supervised learning type. When encountering AI-based predictive models, clinicians and embryologists should be familiar with the machine learning lifecycle (Figure 2):

(1) Collect and pre-process data: Collect relevant data and carry out pre-processing (cleaning, normalising, transforming, etc.) to prepare the data for machine learning algorithms.

(2) Train a machine learning model: Train a machine learning model on the pre-processed data using a suitable algorithm and hyperparameters.

(3) Test and evaluate the model: Test the trained model on a separate test dataset and evaluate its performance using suitable evaluation metrics.

(4) Deploy the model: Deploy the trained model to a production environment, such as a web application or a mobile app.

(5) Monitor the model: Continuously monitor the performance of the deployed model and collect feedback from users.

(6) Refine and update the model: Refine and update the deployed model periodically using new data and feedback to improve its performance and adapt to changing requirements.

Performance evaluation and model validation

When discussing performance, the first step is to define what is being evaluated. If one encounters studies that claim remarkable results on the training dataset, it is advisable to exercise caution. Predicting data that are already in the system makes it easier for the computer to find a previous pattern in the known model, leading to the overfitting effect. It is entirely normal, and almost necessary, for the training set results to be particularly good, as they do not represent the actual predictive potential of the model.

As showed in Figure 3, the process of developing a predictive model involves an initial partition of the test set, which is kept separate from the algorithm's training. Cross-validation is performed on the training set by separating a certain percentage and creating the model with the training set, then predicting the validation set. This process can be repeated several times to obtain cross-validation metrics. This prediction can already be considered representative of the model's predictive potential. Cross-validation can be performed through k-fold cross-validation (e.g., 80% of the dataset for training and 20% for validation) [18,28]; as well as training the model with the full dataset except for one specimen, predicting it, and repeating the process for all specimens in the dataset (leave-one-out cross-validation) [10].

Finally, the test dataset is used to validate how the method (training set+validation set) predicts data that are not in the database. Therefore, it can be considered representative of the model's predictive potential [37].

Performance metrics for machine learning

Depending on the type of algorithm, different metrics should be chosen to evaluate its performance [56]. For regression models, common metrics include mean squared error, mean absolute error, root mean squared error, and r2 [35]. For classification models, common metrics are obtained from a confusion matrix, which unfortunately is not always provided in studies. Common metrics include accuracy, AUC and AUC precision (positive predictive value), recall (sensitivity), negative predictive value and specificity [58]. The F1-score and Matthews correlation coefficient are also metrics to be considered, especially in imbalanced datasets [27]. It is important to ensure that the positive reference is correctly identified in order to avoid confusion when evaluating model performance. For example, in a comparison of euploidy, it may seem obvious that the aneuploid group should be considered as the negative reference. However, the computer

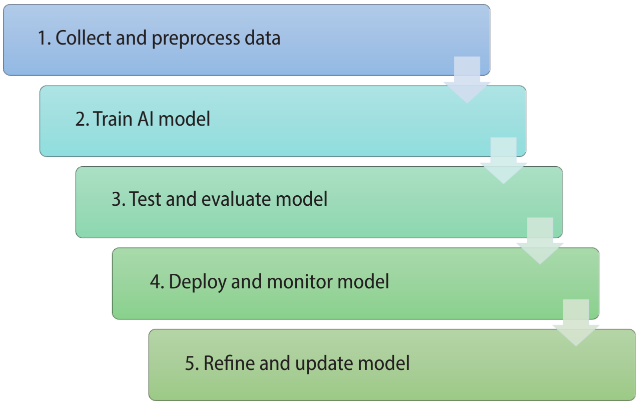
1. Collect and preprocess data
2. Train AI model
3. Test and evaluate model
4. Deploy and monitor model
5. Refine and update model

Figure 2. Machine learning model lifecycle. AI, artificial intelligence.

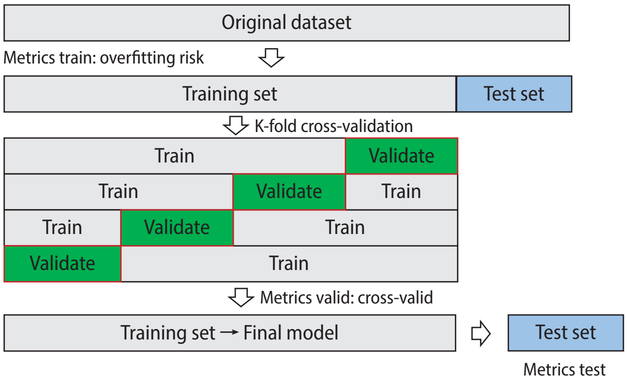

Figure 3. Performance evaluation and model validation using training, validation, and test sets.

may mistakenly assign the aneuploid group as the positive reference if not explicitly specified, such as in cases where alphabetical ordering is used. Therefore, it is crucial to carefully define the positive and negative references before assessing a model's performance.

Conclusions: time to implement?

Different authors have expressed their thoughts on whether or not to implement predictive AI models into the daily practice [59-61]. From my point of view, it is worth considering implementing an algorithm if its result is robust enough to answer the initial question of the requirement. For instance, if the objective was to improve the implantation rate, it is not as crucial whether the embryo selection model is based on viability, genetics, or a combination of both [36,40], nor is the specific value of AUC achieved particularly relevant. While a better AUC is theoretically associated with a better implantation outcome, this cut-off value would not be relevant if the implantation rate with the AI score is superior to that without AI. Nevertheless, external validation should be carried out to verify that the response to the requirement for integrating an AI system in the laboratory is truly satisfactory when applying AI compared to not applying AI. From there, it will be necessary to consider verifying the data either prospectively or in a multi-center setting.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

ORCID

Enric Güell https://orcid.org/0000-0001-6750-1748

References

1. Dimitriadis I, Zaninovic N, Badiola AC, Bormann CL. Artificial intelligence in the embryology laboratory: a review. Reprod Biomed Online 2022;44:435-48.
2. Martin-Climent P, Moreno-Garcia JM. Aplicación de la inteligencia artificial en el laboratorio de reproducción asistida. Trabajo de revisión. Med Reprod Embriol Clin 2022;9:100119.
3. Chow DJ, Wijesinghe P, Dholakia K, Dunning KR. Does artificial intelligence have a role in the IVF clinic? Reprod Fertil 2021;2:C29-34.
4. Fernandez EI, Ferreira AS, Cecilio MH, Cheles DS, de Souza RC, Nogueira MF, et al. Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. J Assist Reprod Genet 2020;37:2359-76.
5. Zaninovic N, Rosenwaks Z. Artificial intelligence in human in vitro fertilization and embryology. Fertil Steril 2020;114:914-20.
6. 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;35:714-25.
7. Blank C, Wildeboer RR, DeCroo I, Tilleman 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; 111:318-26.
8. Liao Q, Zhang Q, Feng X, Huang H, Xu H, Tian B, et al. Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. Commun Biol 2021;4:415.
9. Mendizabal-Ruiz G, Chavez-Badiola A, Aguilar Figueroa I, Martinez Nuno V, Flores-Saiffe Farias A, Valencia-Murilloa R, et al. Computer software (SiD) assisted real-time single sperm selection associated with fertilization and blastocyst formation. Reprod Biomed Online 2022;45:703-11.
10. Zeadna A, Khateeb N, Rokach L, Lior Y, Har-Vardi I, Harlev A, et al. Prediction of sperm extraction in non-obstructive azoospermia patients: a machine-learning perspective. Hum Reprod 2020; 35:1505-14.
11. Nayot D, Mercuri N, Krivoi A, Casper RF, Meriano J, Fjeldstad J. A novel non-invasive oocyte scoring system using AI applied to 2-dimensional images. Fertil Steril 2021;116(3 Suppl):E474.
12. Zhao M, Li H, Li R, Li Y, Luo X, Li TC, et al. Automated and precise recognition of human zygote cytoplasm: a robust image-segmentation system based on a convolutional neural network. Biomed Signal Process Control 2021;67:102551.
13. Kan-Tor Y, Zabari N, Erlich I, Szeskin A, Amitai T, Richter D, et al. Automated evaluation of human embryo blastulation and implantation potential using deep-learning. Adv Intell Syst 2020;2: 2000080.
14. Feyeux M, Reignier A, Mocaer M, Lammers J, Meistermann D, Barriere P, et al. Development of automated annotation software for human embryo morphokinetics. Hum Reprod 2020;35:557-64.
15. Lucio CM, Lopez JT, Zamora ML, Esteve AG, Martinez MB, Suarez ME, et al. CHLOE (FAIRTILITY) can automatically annotate images from time-lapse cultured embryos for Pronucleate (PN) count, morphokinetics and ranking according to ploidy and implantation potential, with a strong agreement compared to experienced embryologists. Reprod Biomed Online 2022;45:e34-5.
16. Khosravi P, Kazemi E, Zhan Q, Toschi M, Malmsten JE, Hickman C, et al. Robust automated assessment of human blastocyst quality using deep learning. bioRxiv 2018 Aug 20 [Priprint]. https://doi.

org/10.1101/394882

17. Yang L, Peavey M, Kaskar K, Chappell N, Devlin DJ, Woodard T, et al. Predicting clinical pregnancy by machine learning algorithm using noninvasive embryo morphokinetics at an academic center. Fertil Steril 2019;112:e181.
18. Liang B, Gao Y, Xu J, Song Y, Xuan L, Shi T, et al. Raman profiling of embryo culture medium to identify aneuploid and euploid embryos. Fertil Steril 2019;111:753-62.
19. Chavez-Badiola A, Flores-Saiffe-Farias A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reprod Biomed Online 2020;41:585-93.
20. Diakiw SM, Hall JM, VerMilyea MD, Amin J, Aizpurua J, Giardini L, et al. Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. Hum Reprod 2022;37:1746-59.
21. Diakiw SM, Hall JM, VerMilyea M, Lim AY, Quangkananurug W, Chanchamroen S, et al. An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos. Reprod Biomed Online 2022;45:1105-17.
22. Barnes J, Brendel M, Gao VR, Rajendran S, Kim J, Li Q, et al. A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study. Lancet Digit Health 2023;5:e28-40.
23. Berntsen J, Rimestad J, Lassen JT, Tran D, Kragh MF. Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. PLoS One 2022;17:e0262661.
24. Theilgaard Lassen J, Fly Kragh M, Rimestad J, Nygard Johansen M, Berntsen J. Development and validation of deep learning based embryo selection across multiple days of transfer. Sci Rep 2023;13:4235.
25. Glatstein I, Chavez-Badiola A, Curchoe CL. New frontiers in embryo selection. J Assist Reprod Genet 2023;40:223-34.
26. Cimadomo D, Marconetto A, Trio S, Chiappetta V, Innocenti F, Albricci L, et al. Human blastocyst spontaneous collapse is associated with worse morphological quality and higher degeneration and aneuploidy rates: a comprehensive analysis standardized through artificial intelligence. Hum Reprod 2022;37:2291-306.
27. De Gheselle S, Jacques C, Chambost J, Blank C, Declerck K, De Croo I, et al. Machine learning for prediction of euploidy in human embryos: in search of the best-performing model and predictive features. Fertil Steril 2022;117:738-46.
28. Ortiz JA, Morales R, Lledo B, Vicente JA, Gonzalez J, Garcia-Hernandez EM, et al. Application of machine learning to predict aneuploidy and mosaicism in embryos from in vitro fertilization cycles. AJOG Glob Rep 2022;2:100103.
29. Yu JL, Su YF, Zhang C, Jin L, Lin XH, Chen LT, et al. Tracking of menstrual cycles and prediction of the fertile window via measurements of basal body temperature and heart rate as well as machine-learning algorithms. Reprod Biol Endocrinol 2022;20:118.
30. Bormann CL, Curchoe CL, Thirumalaraju P, Kanakasabapathy MK, Gupta R, Pooniwala R, et al. Deep learning early warning system for embryo culture conditions and embryologist performance in the ART laboratory. J Assist Reprod Genet 2021;38:1641-6.
31. Khodabandelu S, Basirat Z, Khaleghi S, Khafri S, Montazery Kordy H, Golsorkhtabaramiri M. Developing machine learning-based models to predict intrauterine insemination (IUI) success by address modeling challenges in imbalanced data and providing modification solutions for them. BMC Med Inform Decis Mak 2022;22:228.
32. Letterie G, Mac Donald A. Artificial intelligence in in vitro fertilization: a computer decision support system for day-to-day management of ovarian stimulation during in vitro fertilization. Fertil Steril 2020;114:1026-31.
33. Hariton E, Chi EA, Chi G, Morris JR, Braatz J, Rajpurkar P, et al. A machine learning algorithm can optimize the day of trigger to improve in vitro fertilization outcomes. Fertil Steril 2021;116:1227-35.
34. Fanton M, Nutting V, Solano F, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for predicting the optimal day of trigger during ovarian stimulation. Fertil Steril 2022;118:101-8.
35. Correa N, Cerquides J, Arcos JL, Vassena R. Supporting first FSH dosage for ovarian stimulation with machine learning. Reprod Biomed Online 2022;45:1039-45.
36. Guell Penas E, Vives Perello A, Esquerra Pares M, Mladenova Koleva M. P-179 Aneuploid embryos as a proposal for improving artificial intelligence performance. Hum Reprod 2022;37(Supplement 1):deac106-059.
37. Goyal A, Kuchana M, Ayyagari KP. Machine learning predicts livebirth occurrence before in-vitro fertilization treatment. Sci Rep 2020;10:20925.
38. Miyagi Y, Habara T, Hirata R, Hayashi N. Feasibility of deep learning for predicting live birth from a blastocyst image in patients classified by age. Reprod Med Biol 2019;18:190-203.
39. Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can predict embryo ploidy status based on time-lapse data. Reprod Biol Endocrinol 2021;19:185.
40. Loewke K, Cho JH, Brumar CD, Maeder-York P, Barash O, Malmsten JE, et al. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. Fertil Steril

2022;117:528-35.

41. Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Hum Reprod 2019;34:1011-8.

42. VerMilyea M, Hall JM, Diakiw SM, Johnston A, Nguyen T, Perugini D, 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.

43. Cimadomo D, Chiappetta V, Innocenti F, Saturno G, Taggi M, Marconetto A, et al. Towards automation in IVF: pre-clinical validation of a deep learning-based embryo grading system during PGT-A cycles. J Clin Med 2023;12:1806.

44. Afnan MAM, Rudin C, Conitzer V, Savulescu J, Mishra A, Liu Y, et al. Ethical implementation of artificial intelligence to select embryos in in vitro fertilization. In: AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; 2021 May 19-21; Virtual Event, USA. Association for Computing Machinery. 2021; pp 316-26. Available from: https://doi.org/10.1145/3461702.3462589

45. Martinez-Granados L, Serrano M, Gonzalez-Utor A, Ortiz N, Badajoz V, Olaya E, et al. Inter-laboratory agreement on embryo classification and clinical decision: conventional morphological assessment vs. time lapse. PLoS One 2017;12:e0183328.

46. Fordham DE, Rosentraub D, Polsky AL, Aviram T, Wolf Y, Perl O, et al. Embryologist agreement when assessing blastocyst implantation probability: is data-driven prediction the solution to embryo assessment subjectivity? Hum Reprod 2022;37:2275-90.

47. Barrie A, McDowell G, Troup S. An investigation into the effect of potential confounding patient and treatment parameters on human embryo morphokinetics. Fertil Steril 2021;115:1014-22.

48. Meseguer M, Valera MA. The journey toward personalized embryo selection algorithms. Fertil Steril 2021;115:898-9.

49. Curchoe CL, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. Fertil Steril 2020;114:921-6.

50. Zhu X, Wu X. Class noise vs. attribute noise: a quantitative study. Artif Intell Rev 2004;22:177-210.

51. Frenay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 2014;25:845-69.

52. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst 2023;34:8135-53.

53. Popovic M, Dheedene A, Christodoulou C, Taelman J, Dhaenens L, Van Nieuwerburgh F, et al. Chromosomal mosaicism in human blastocysts: the ultimate challenge of preimplantation genetic testing? Hum Reprod 2018;33:1342-54.

54. Orvieto R, Jonish-Grossman A, Maydan SA, Noach-Hirsh M, Dratviman-Storobinsky O, Aizer A. Cleavage-stage human embryo arrest, is it embryo genetic composition or others? Reprod Biol Endocrinol 2022;20:52.

55. Kragh MF, Karstoft H. Embryo selection with artificial intelligence: how to evaluate and compare methods? J Assist Reprod Genet 2021;38:1675-89.

56. Nasir V, Sassani F. A review on deep learning in machining and tool monitoring: methods, opportunities, and challenges. Int J Adv Manuf Technol 2021;115:2683-709.

57. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. Stat Surv 2022;16:1-85.

58. Swain J, VerMilyea MT, Meseguer M, Ezcurra D; Fertility AI Forum Group. AI in the treatment of fertility: key considerations. J Assist Reprod Genet 2020;37:2817-24.

59. Fitz VW, Kanakasabapathy MK, Thirumalaraju P, Kandula H, Ramirez LB, Boehnlein L, et al. Should there be an "AI" in TEAM?: embryologists selection of high implantation potential embryos improves with the aid of an artificial intelligence algorithm. J Assist Reprod Genet 2021;38:2663-70.

60. Simopoulou M, Sfakianoudis K, Maziotis E, Antoniou N, Rapani A, Anifandis G, et al. Are computational applications the "crystal ball" in the IVF laboratory?: the evolution from mathematics to artificial intelligence. J Assist Reprod Genet 2018;35:1545-57.

61. Sfakianoudis K, Maziotis E, Grigoriadis S, Pantou A, Kokkini G, Trypidi A, et al. Reporting on the value of artificial intelligence in predicting the optimal embryo for transfer: a systematic review including data synthesis. Biomedicines 2022;10:697.