Human Reproduction, Vol.36, No.9, pp. 2429–2442, 2021
Advance Access Publication on July 29, 2021 doi:10.1093/humrep/deab168

human reproduction

MINI REVIEW

Artificial intelligence in the fertility clinic: status, pitfalls and possibilities

M.A. Riegler [1,*], M.H. Stensen[2], O. Witczak[3], J.M. Andersen[3], S.A. Hicks[1,4], H.L. Hammer[1,4], E. Delbarre [3], P. Halvorsen[1,4], A. Yazidi[4], N. Holst[2], and T.B. Haugen[3]

[1]Department of Holistic Systems, Simula Metropolitan Center for Digital Engineering, Oslo, Norway [2]Fertilitetssenteret, Oslo, Norway [3]Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet—Oslo Metropolitan University, Oslo, Norway [4]Department of Computer Science, Faculty of Technology, Art and Design, OsloMet—Oslo Metropolitan University, Oslo, Norway

*Correspondence address. Department of Holistic Systems, Simula Metropolitan Center for Digital Engineering, Oslo 0167, Norway. E-mail: michael@simula.no https://orcid.org/0000-0002-3153-2064

Submitted on May 19, 2021; resubmitted on June 21, 2021; editorial decision on June 23, 2021

ABSTRACT: In recent years, the amount of data produced in the field of ART has increased exponentially. The diversity of data is large, ranging from videos to tabular data. At the same time, artificial intelligence (AI) is progressively used in medical practice and may become a promising tool to improve success rates with ART. AI models may compensate for the lack of objectivity in several critical procedures in fertility clinics, especially embryo and sperm assessments. Various models have been developed, and even though several of them show promising performance, there are still many challenges to overcome. In this review, we present recent research on AI in the context of ART. We discuss the strengths and weaknesses of the presented methods, especially regarding clinical relevance. We also address the pitfalls hampering successful use of AI in the clinic and discuss future possibilities and important aspects to make AI truly useful for ART.

Key words: artificial intelligence / machine learning / ART / embryology / semen analysis / embryo / spermatozoa / fertility / infertility / algorithm

Introduction

The number of treatments with ART is steadily increasing in Europe, and in 2016, over 900 000 treatment cycles were performed (Wyns *et al.*, 2020). Even though there have been gradual improvements in the success rate, only one-third of the ART cycles result in a live birth, and only 5% of the aspirated oocytes have the competence to develop into a child (Lemmen *et al.*, 2016; Wyns *et al.*, 2020). This implies that there is potential for improvement in the crucial steps in ART treatments, such as the selection of embryos for transfer and the selection of spermatozoa for ICSI. Improving the ability to select a single embryo with the highest implantation potential could increase live birth rates and time to pregnancy, as well as minimise the chance of multiple pregnancies due to the transfer of multiple embryos. Likewise, a more reliable method for sperm selection may increase the success rates of the ICSI procedure. Furthermore, the disputable clinical value of semen analysis in male fertility investigation and for ART justifies a need for improving the methods of sperm evaluation both for diagnostic purposes and for decisions regarding the fertilisation method of the ART treatment.

Video and image analysis constitutes a major part of ART, and artificial intelligence (AI) methods are especially suited for image classification. In addition to videos and images, AI can be used to analyse other types of data, like text or tabular data. As in other parts of medicine, AI methods have been introduced in the field of ART. They have the advantage of objectivity and have the potential to improve ART, which in some parts are based on subjective assessments.

In this review, we provide an overview of studies found in Embase (Ovid), where AI methods have been applied in human reproductive medicine with an emphasis on ART. Furthermore, we discuss how to avoid the pitfalls and describe the potential use of AI in clinical practice in the future.

Current challenges in ART

Highly trained personnel in fertility clinics are faced with important and difficult decisions every day, such as deciding which fertilisation method to use, which spermatozoon to select for ICSI, and which embryo to transfer to the uterus. One of the major challenges in the subjective assessments of embryos is the high intra- and inter-operator variability which exists in the evaluation of morphology and morphokinetics (Paternot *et al.*, 2009; Sundvall *et al.*, 2013; Storr *et al.*, 2017). With time-lapse technology, embryos can be monitored continuously, and

the complete process of embryo development is more precisely assessed. However, there is no evidence that the use of time-lapse technology has improved live birth rates after ART (Armstrong et al., 2019).

Whilst sperm morphology has no definite impact on the outcome after ART, sperm concentration and sperm motility are normally assessed for deciding whether IVF or ICSI should be used as the fertilisation method (Høst et al., 2001). Strikingly, ICSI is increasingly used irrespective of a male factor infertility diagnosis (Boulet et al., 2015; Vander Borght and Wyns, 2018). Among the cycles reported in Europe in 2016, 28% were IVF and 72% ICSI (Wyns et al., 2020), although the male factor accounts for only 20–30% of the diagnoses of the infertile couples. This is of increasing concern since performing ICSI instead of IVF in couples where the male partner has a defined normal semen sample does not increase the live birth rate (Dang et al., 2021).

Early in the fertility investigation, a standard semen analysis according to WHO guidelines (WHO, 2010) is usually performed. This analysis might reveal information essential for deciding whether ART should be recommended as a treatment. The method is time-consuming and prone to limited reproducibility and high inter-personnel variation (Tomlinson, 2016). Several computer-aided sperm analyses (CASA) systems are available, but they are still most suitable for assessing spermatozoa separated from seminal plasma, and their reliability is debatable (Mortimer et al., 2015).

When selecting spermatozoa to inject for ICSI, the procedure is performed by visually evaluating the morphology and motility of spermatozoa with an ICSI microscope. This selection process is subjective, based on a qualitative evaluation of the operator, and not on objective sperm characteristics.

The potential of AI in ART

New technologies, such as better cameras and data capturing systems, are rapidly becoming an integrated part of the fertility clinic and result in a vast amount of stored data, including patient data, embryo time-lapse videos and sperm videos. In recent years, AI has proved to be a valuable tool in medicine by analysing large amounts of data (Hosny et al., 2018; Yang and Bang, 2019). A typical approach for using AI models in ART can be seen in Fig. 1. In particular, machine learning (ML), a subfield within AI, refers to algorithms that automatically learn from data without being explicitly programmed.

An overview of common AI methods used in ART is given in Fig. 2. Supervised and unsupervised learning are subgroups of ML. Supervised learning refers to methods that learn from datasets where the answer (the label) is given for each observation. An observation within a dataset could be data from an ART cycle, like an embryo image, and the label regarding whether the embryo resulted in a pregnancy or not. The algorithm will learn from the dataset, and the resulting ML model can be used to predict pregnancy or not for data from another ART cycle with unknown labels. Unsupervised learning refers to methods that search for patterns in unlabelled data, for example, automatically grouping blastocyst images based on visual features automatically determined by the algorithm that may correlate with morphological characteristics. Such visual features can be completely different from what human observers are able to recognise or may see as relevant. Artificial neural networks (ANNs) are a class of supervised learning, and deep neural networks (DNNs), or deep learning (DL), refers to especially large and complex ANNs. DL methods have the ability to learn from unstructured data such as images or text.

Details of studies discussed in this review can be found in Table I for embryo related articles and in Table II for sperm related articles.

AI in embryo assessment

Most articles about embryo assessment and selection for transfer address the prediction of embryo quality, grading and ranking, and compare the performance of the AI model with an evaluation done by embryologists (Dirvanauskas et al., 2019; Kanakasabapathy et al., 2019; Khosravi et al., 2019; Raudonis et al., 2019; Fukunaga et al., 2020; Bormann et al., 2020a, 2020b; Rad et al., 2020; Zhao et al., 2021). To make an automatic grading system, the model must learn to locate the embryo in the dish, segment important features, and then assess and grade the embryo from manually annotated data. Manual annotations provided by embryologists are time-consuming to create, leading to small and sparsely annotated datasets. Therefore, most studies of AI methods and resulting models in ART can be considered preliminary. With the development of time-lapse technology, access to image and video data has become more available, making it possible to utilise this data to build new AI models. Dirvanauskas et al. (2019) predicted embryo development stages by time-lapse videos using features extracted from a Convolutional Neural Network (CNN). In one study, an automated system was established to detect pronuclei in time-lapse images with the precision almost equivalent to highly skilled embryologists (Fukunaga et al., 2020). In another study, the zona pellucida (ZP) and the cytoplasm and pronucleus in zygotes were detected by developing an algorithm using DL image segmentation technology (Zhao et al., 2021). One group reported the possibility of identifying human embryo development stages (Raudonis et al., 2019). First, the location of an embryo in the image was detected by employing a visual image feature-based classifier. Then, a multi-class prediction model was developed to predict the cell stage of the embryo using DL. Others reported a system to detect and assess blastocyst quality by using DL to detect the ZP area (Rad et al., 2018).

Data augmentation techniques, like cropping and resizing which are usually used to increase dataset size or variation, were applied to embryo assessment to compensate for the lack of data for training the DL models (Rad et al., 2020). Augmented images were proven to be effective in filling the generalisation gap when available data is limited. Experimental results confirmed that the proposed models were capable of segmenting trophectoderm (TE) regions.

Inner cell mass (ICM) has been assessed by a computer-based and semi-automatic grading of human blastocysts (Santos Filho et al., 2012). A CNN was able to predict ICM and TE grades from a single frame (a frame is an image extracted from a video), and a recurrent neural network was applied on top to incorporate temporal information of the expanding blastocysts from multiple frames. Additionally, when evaluating implantation rates for embryos grouped by morphology grades, a CNN provided a slightly higher correlation between predicted embryo quality and implantation ability than did human

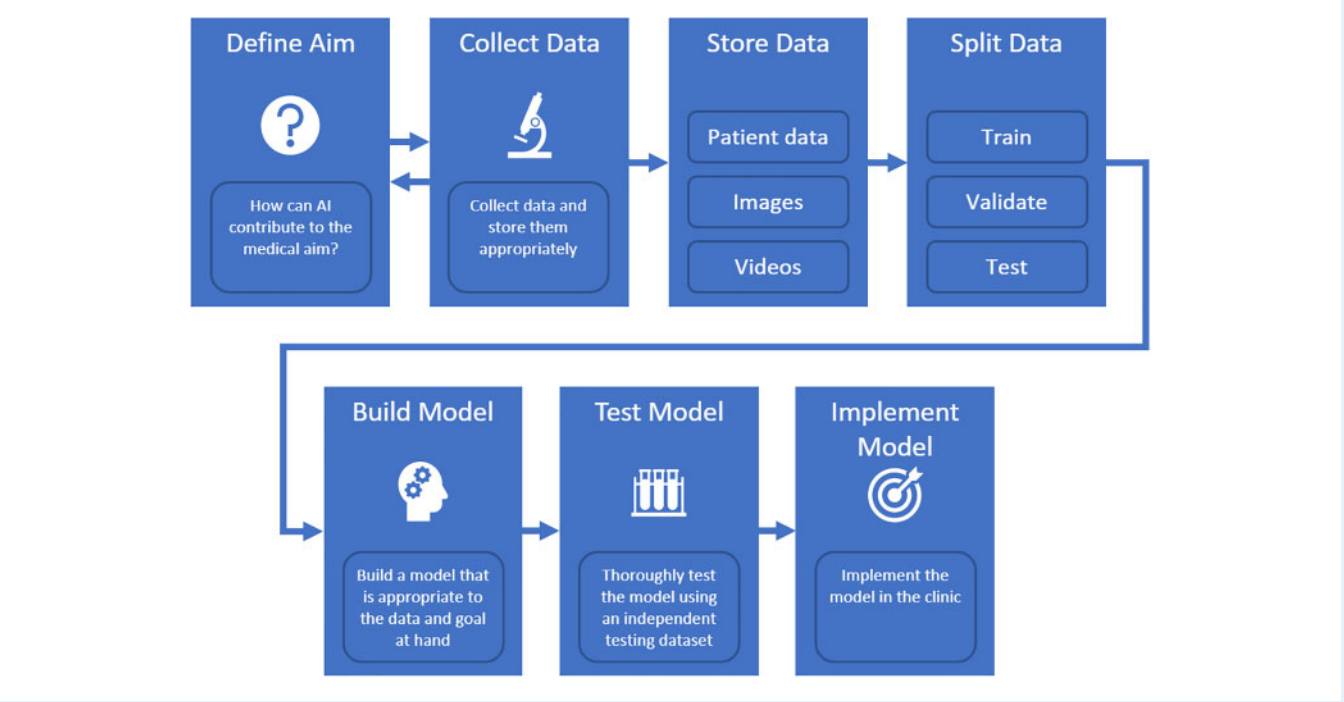

Figure 1. Development of a machine learning model. To implement a machine learning model at the clinic, at first a clinically relevant aim should be defined, and data must be collected in line with this aim. The collected data should then be stored in an appropriate format so that the machine learning algorithm can process it. The stored data should be split into a training, validation and testing partitions to ensure a robust and thorough evaluation. In the optimal case the testing dataset is provided from an independent source (different clinic, new patients). These parts are then be used to build a model that is in line with the medical goal. After the model is built, it should be thoroughly evaluated to verify its generalisability and to avoid unintended biases. Once the model has been thoroughly tested, it can be implemented in the clinic. The model should be continuously monitored and tested while in production and as the circumstances required are updated.

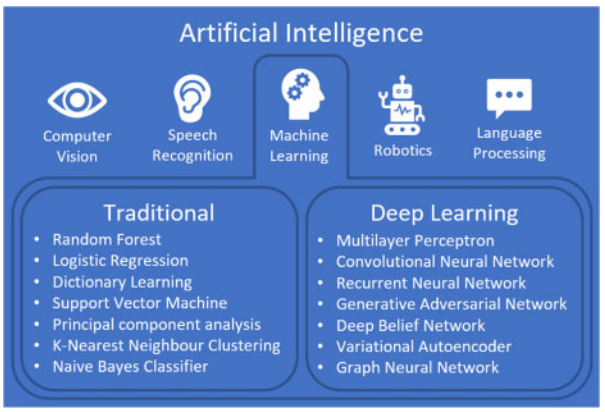

Figure 2. Subfields defined by artificial intelligence. Machine learning is the most relevant field for the current development of artificial intelligence system for the clinic. Machine learning can further be split into traditional machine learning methods and deep learning. Note that the subfields are not mutually exclusive; most of them rely heavily on machine learning, like computer vision and language processing.

embryologists (Kragh *et al.*, 2019). The use of a CNN trained to assess an embryo's implantation potential directly, when using euploid embryos capable of implantation, outperformed 15 trained embryologists (Bormann *et al.*, 2020a).

In a retrospective analysis of time-lapse videos and clinical outcomes of 10 000 embryos from eight different IVF clinics across four different countries, a DL model was built with a high level of predictability regarding the embryo implantation likelihood (Tran *et al.*, 2019). A prospective double-blinded study using retrospective data addressed the variability between embryologists to select embryos for biopsy and cryopreservation (Bormann *et al.*, 2020b). It was found that the application of a DNN could improve the reliability and perform with high consistency during the process of embryo selection, thereby potentially improving outcomes.

A DL-based system called Life Whisperer showed a sensitivity of 70% for viable embryos while maintaining a specificity of 61% for non-viable embryos across three independent blind test sets from different clinics (Ver Milyea *et al.*, 2020). The model demonstrated a 25% increase over embryologists for accuracy, and the ranking comparison demonstrated an improvement of 42% over embryologists. One embryo ranking model increased the success of ART treatments in oocyte donation programs (Alegre *et al.*, 2021). The

Table I Overview of studies using AI-methods in embryo assessment and selection, and for prediction before treatment.

Year	Study	Aim of the study	Outcome	Dataset	AI methods	Summary answer
2017	Milewski et al.	Investigating the potential of using data on embryo implantation and morphokinetic parameters in predictive AI models.	Probability of implantation, clinical pregnancy.	A dataset of time-lapse recordings of 610 embryos from 514 treatment cycles, morphokinetic parameters, data on implantation, women's age. It is unclear if the dataset was prospectively collected.	Traditional ML (Principal Component Analysis) and Deep Learning (Multilayer Perceptron)	Morphokinetic parameters from the time-lapse videos used to discriminate between implanted and nonimplanted achieved an AUC of 0.71.
2018	Rad et al.	Automatic segmentation of the Zona Pellucida.	Segmentation of Zona Pellucida.	A retrospective dataset consisting of images of blastocyst.	Deep Learning	The AI model was able to segment the Zona Pellucida with an IoU score of 0.78.
2019	Tran et al.	Predict the probability of pregnancy with foetal heart from time-lapse videos.	Foetal heart pregnancy or not.	A retrospective dataset containing time-lapse videos of 10,638 embryos cultured to blastocyst stage from 1,648 patients across 8 IVF clinics. No manual assessment of videos.	Deep Learning (CNN)	AI model (IVY) was able to predict the probability of fetal heart pregnancy based on timelapse videos with a mean AUC of 0.93.
2019	Dirvanauskas et al.	Predict embryo development stage from time-lapse videos.	Embryo development stage (1-cell, 2-cell, 4-cell, 8-cell, no embryo).	A retrospective dataset containing 7,002 time-lapse images from 10 embryos.	Deep learning (CNN) and traditional ML (K Nearest Neighbour, Cecoc, Decision Trees, Naive Bayes Classifier)	The AI model for embryo classification achieved an accuracy of 97.62%.
2019	Kanakasabapathy et al.	Develop inexpensive platforms for use in a stand-alone optical system and a smartphone-based optical system for automated grading of embryos based on images.	Classification of embryos based on cell morphology.	A retrospective dataset containing 160 embryo images from a stand-alone optical system and 385 embryo images from a smartphone-based optical system. Models were pretrained on other high-quality embryo data.	Deep Learning (CNN)	Two systems were developed for grading embryos (stand-alone imaging system and smartphone optical system). Both systems achieve an accuracy above 90%.
2019	Khosravi et al.	Develop an AI model for accurate prediction of blastocyst quality and selection for single embryo for transfer.	Classification of embryos into poor-quality and good-quality.	A retrospective dataset containing 12,001 time-lapse images at 110 hr post-insemination from 10,148 embryos. Manual classification by embryologists. Age of patient was included in the model for 2,182 embryos. Two external datasets were used for validation.	Deep Learning (CNN)	AI model (STORK) predicted blastocyst quality with an AUC above 0.98. The model achieved an AUC of 0.90 and 0.76 respectively on two datasets from other clinics.
2019	Kragh et al.	Develop AI method for automatic grading of blastocyst morphological appearance based on time-lapse images.	Inner cell mass and trophectoderm grading, implantation rate.	A dataset containing time-lapse videos of 4,483 embryos (both IVF and ICSI treatment). All images were graded by embryologists. Implantation information for 287 embryos. It is unclear if the dataset was prospectively collected.	Deep Learning (CNN, Recurrent Neural Network)	AI model achieved an accuracy of 65% for inner cell mass grading and 70% for trophectoderm grading. Prediction of implantation achieved an AUC of 0.66.

(continued)

Table I Continued

Year	Study	Aim of the study	Outcome	Dataset	AI methods	Summary answer
2019	Raudonis *et al.*	Automatically detect human embryo development stages during incubation.	Detect embryo in an image and classify the embryo development stage (1-cell, 2-cell, 3-cell, 4-cell, > 4-cell).	A dataset containing images of early-stage embryo development from an ESCO Miri TL incubator system. It is unclear if the dataset was prospectively collected.	Deep Learning (CNN)	Two AI models were considered, both achieved a stage classification accuracy above 92%. The most difficult stage to classify was 3-cell.
2019	Qiu *et al.*	Prediction of a clinical model for estimating the cumulative live birth chance of the first complete IVF cycle using pretreatment variables including BMI and AMH.	Cumulative live birth chance before IVF.	A retrospective dataset containing age, AMH, BMI, duration of infertility, previous live birth, previous miscarriage, previous abortion, and type of infertility.	Traditional ML (Logistic Regression, Random Forest, XGBoost, Support Vector Machine)	Four machine learning models were tested, of which XGBoost achieved the best score with an AUC of 0.73. The results indicate that BMI and AMH have a significant impact on live birth.
2019	Vogiatzi *et al.*	Predict live birth from embryo variables by including parameters that exert a meaningful effect on live birth following assisted reproduction.	Live birth or not.	12 input features: Age (female), Age at menarche, Difficulty during ET, Endometrium thickness prior to OR, ET/2PN, TQE D3, TQE D3/2PN, Total gonadotropins, Age group, Dyspareunia, Fresh or frozen cycle, Menarche > 12 years. The dataset was collected retrospectively.	Deep Learning (Multilayer Perceptron)	A multilayer perceptron using the 12 input features achieved a sensitivity of 0.71 and a specificity of 0.70 for predicting live birth.
2020	Bori *et al.*	Describe novel embryo features for implantation potential prediction that may be used as input data in AI models.	Prediction of implantation potential.	A retrospective dataset containing time-lapse images from 637 embryos (ICSI-cycles without PGT, single fresh embryo transfer), Implantation rate based on foetal heartbeat ultrasound after eight weeks. Oocyte donation programme.	Deep Learning (Multilayer Perceptron)	Two novel embryo features with significantly different values in implanted and nonimplanted embryos were identified. Novel embryo features, in addition to conventional morphokinetic parameters, can improve predictive AI models.
2020	Bormann *et al.* (a)	Evaluation of AI models for embryo selection based on images.	Embryo quality and implantation potential.	A retrospective dataset containing single timepoint images at 113 h post-insemination for 742 embryos from 97 patients.	Deep Learning (CNN)	Two AI models were evaluated. One selected the highest quality embryo with 90% accuracy, and the other was able to assess implantation potential better than trained embryologists from different fertility centres.
2020	Bormann *et al.* (b)	Evaluate AI models for embryo quality scoring and assessment of biopsy or cryopreservation of blastocysts, compared to decisions by trained embryologists.	Morphological quality on a 1–5 scale.	For embryo scoring, images from 3469 embryos. 748 at 70 h post-insemination and 742 images at 113 h post-insemination. For biopsy and cryopreservation assessment, 56 blastocysts images at 113 h post-insemination. All images were evaluated by	Deep Learning (CNN)	The AI models showed less variability in embryo grading than embryologists and outperformed the embryologists in selecting blastocyst biopsy and cryopreservation.

(continued)

Table I Continued

Year	Study	Aim of the study	Outcome	Dataset	AI methods	Summary answer
				trained embryologists. Both datasets were retrospectively collected.		
2020	Chavez-Badiola et al. (a)	Evaluate AI model performance for prediction of ploidy and implantation compared to trained embryologists.	Embryo ranking, embryo ploidy.	A retrospective dataset containing single time-point images from 840 embryos at day 5 or 6 after fertilization by ICSI. Ploidy, hCG results, or both were known.	Deep Learning (Multilayer Perceptron)	An AI model (ERICA) was able to identify and rank blastocysts with the best potential from one image with higher accuracy than embryologists.
2020	Chavez-Badiola et al. (b)	Predict pregnancy test results after embryo transfer.	Successful pregnancy or not.	A retrospective dataset containing embryo images and patient age.	Traditional ML (Probabilistic Bayesian, Support Vector Machine, Decision Trees, Random Forest) and Deep Learning (Multilayer Perceptron)	Several AI models were tested, of which the support vector machine achieved the best result across three datasets.
2020	Fukunaga et al.	Automatic pronuclei counting using deep learning.	Number of pronuclei.	A dataset containing 900 time-lapse images of 300 embryos up to 20 h post-insemination. 70 images of each embryo. Manual assessment and annotation of pronuclei. It is unclear if the dataset was prospectively collected.	Deep Learning (CNN)	The AI model was able to count pronuclei with a sensitivity of 99% for 0PN, 82% for 1PN, and 99% for 2PN. The system performed similarly to that of trained human experts.
2020	Rad et al.	Automatic trophectoderm segmentation in human embryo using deep learning.	Trophectoderm segmentation.	A retrospective dataset containing images of day-5 human embryo.	Deep Learning (CNN, Generative Adversarial Networks)	An AI model was used to segment human embryos. The model achieved an IoU score of 76.71.
2020	Raef et al.	Predict implantation outcome after embryo transfer cycle.	Implantation rate. Positive or negative beta-HCG.	A dataset containing 82 features (patient-related data, female and male pathology, semen analysis, lab tests, oocyte and embryo data and PRP) Attributes related to implantation arranged in two groups (N = 82): 1) patient-related features (N = 59) and 2) ART cycle features (N = 23). It is unclear if the dataset was prospectively collected.	Traditional ML (Naive Bayes Classifier, Support Vector Machine, Random Forest, K Nearest Neighbour, Decision Trees) and Deep Learning (Multilayer Perceptron)	Six AI models were tested, where the random forest algorithm achieved the best result with an accuracy of 90.4% and an AUC of 93.7%.
2020	Ver Milyea et al.	Predict embryo viability using images captured by optical light microscopy.	Implantation rate—foetal heartbeat.	A retrospective dataset containing light microscopy images of blastocysts, clinical outcome.	Deep Learning (Convolutional Neural Network)	An AI model (Life Whisperer) was tested on three independent testing datasets, where it achieved a 70.1% sensitivity for viable embryos and a specificity of 60.5% for non-viable embryos.
2020	Goyal et al.	Predict live birth before IVF treatment.	Live birth or not.	A retrospective dataset containing 141,160 patient records,	Deep Learning (Multilayer Perceptron)	Several machine learning models were evaluated, of which the multilayer

(continued)

Table I Continued

Year	Study	Aim of the study	Outcome	Dataset	AI methods	Summary answer
				anonymized register data collected from the year 2010–2016 obtained from the Human Fertilisation & Embryology Authority.		perceptron performed best with an F1-Score of 72.94%.
2021	Alegre *et al.*	Evaluate and test an automatic software for embryo evaluation and selection (Dana).	Embryo implantation potential.	A retrospective dataset containing time-lapse images and patient characteristics from oocyte donation program. Phase 1: 1,676 embryos from 955 couples. Phase 2: 996 embryos from 249 cycles (multiple centres). Phase 3 147 embryos from 108 patients.	Deep Learning (CNN)	Increased success of IVF treatment was found with the assistance of automated embryo ranking by Dana. The creation of a data cloud can improve the system further.
2021	Bori *et al.*	Develop an AI model for prediction of live birth based on blastocyst morphology and proteomic profile of culture media.	Prediction of live birth.	A retrospective dataset containing single time point images at 111 hr +/- 1.5 hr from 212 patients. 186 embryos after exclusions (131 non PGT from oocyte donation programme, 55 PDG with proteomic profile.	Deep Learning (Multilayer Perceptron)	Three AI models using both morphological and proteomic variables. The best model predicted live birth with an AUC of 1.0.
2021	Zhao *et al.*	Automatic segmentation of day one embryos in zona pellucida (ZP), cytoplasm, and pronucleus (PN).	Cytoplasm, ZP and PN segments.	A dataset containing images of day-one embryos (zygotes). It is unclear if the dataset was prospectively collected.	Deep Learning (CNN, Generative Adversarial Networks)	The AI model achieved a precision of 97% when segmenting the cytoplasm, 80% for the zona pellucida, and 84% for the pronucleus.

AI, Artificial intelligence; CNN, Convolutional neural network; AUC, Area under the curve; IVF, In vitro fertilization; ICSI, Intracytoplasmic sperm injection; ZP, Zona pellucida; PN, Pronucleus, PGT, Preimplantation genetic testing; AMH, Anti-Mullerian hormone; BMI, body mass index.

multicentre nature of the above study supported its applicability at different clinics, standardising the interpretation of embryo development.

Embryo assessment, ranking, and selection are procedures normally based on evaluations at different time points during embryo development and in several focal planes to get a view of the whole embryo. There are numerous studies where only static images, usually in one single focal plane, are used for the AI analysis, which do not mirror the clinical practice (Rad *et al.*, 2018; Kanakasabapathy *et al.*, 2019; Khosravi *et al.*, 2019; Bormann *et al.*, 2020a, 2020b; Chavez-Badiola *et al.*, 2020a; Chavez-Badiola *et al.*, 2020b; Bori *et al.*, 2021). In these models, well-curated, high-quality data is crucial. For example, non-selection of a large number of images representative of the diversity, inconsistent image treatment or inaccurate labelling of images can lead to poor performing models (Tsipras *et al.*, 2020). Models involving time-lapse videos might also raise problems since the definition of the important morphokinetic markers may vary between different laboratories and still requires an automated and unbiased process (Milewski *et al.*, 2017; Dirvanauskas *et al.*, 2019; Tran *et al.*, 2019; Bori *et al.*, 2020; Alegre *et al.*, 2021).

AI methods should incorporate patient data that may impact the outcome, such as maternal age. A framework (STORK) based on a large collection of human embryo time-lapse images used a CNN to automatically predict blastocyst quality depending on patient age (Khosravi *et al.*, 2019). Milewski *et al.* (2017) extracted several time points and specific relative cleavage times together with fragmentation levels, presence of multinucleation, evenness of blastomeres and woman's age. An ANN was trained to predict embryo implantation from the extracted features. Another study that included 82 features of patient data found that follicle stimulating hormone/human menopausal gonadotropin dosage was the strongest predictor of embryo implantation (Raef *et al.*, 2020).

Table II Overview of studies using AI-methods in semen analysis and selection of sperm for ICSI.

Year	Study	Aim of study	Outcome	Dataset	AI Methods	Summary answer
2014	Chang et al.	Improve AI models for detection of human sperm head characteristics including, acrosome and nucleus.	Sperm morphology	A prospective dataset containing 20 images with a total of 210 stained sperm cells. Sperm cell details were manually classified and annotated in the dataset.	Traditional ML (Clustering)	Models showed 80% overlap with manual classification and more precise sperm head detection and segmentation than previously described models.
2017	Chang et al.	Explore AI modes to classify sperm head morphology into five classes (normal, tapered, pyriform, small, amorphous) and introduce a new dataset.	Sperm morphology	A retrospective dataset containing images of 1,854 stained sperm heads from six semen smears (SCIAN MorphoSpermGS). Sperm head shape was manually classified and annotated in the dataset.	432	The best model was able to obtain 49% correct classification of head shape into the five classes.
2017	Shaker et al.	Explore Dictionary Learning technique for classification of sperm head shapes into four classes (normal, tapered pyriform and amorphous), and introduce a new dataset.	Sperm morphology	Two retrospective datasets. 216 images of stained sperm heads (HuSHeM dataset). Sperm head shape was manually classified and annotated in the dataset. 1133 images from the SCIAN-MorphoSpermGS dataset.	Traditional ML (Dictionary Learning)	Use of Dictionary Learning was more effective for sperm head classification than previously published shape-based features.
2017	Goodson et al.	Development of AI model for classification of sperm motility patterns during invitro capacitation.	Sperm motility	CASA tracks of 2,817 washed sperm cells from 18 subjects. All tacks were manually classified as progressive, intermediate, hyperactivated, slow, weakly motile. It is unclear if the dataset was prospectively collected.	Traditional ML (Support Vector Machine, Decision Tree)	A web-based program, CASAnova, was developed. This program classifies sperm motility patterns into one of five classes with an overall accuracy of 89.9%.
2019	Agarwal et al.	Evaluate the performance of an automated AI system (LensHook) to measure sperm concentration and sperm motility.	Sperm concentration and sperm motility	A prospective dataset containing images and video from 135 semen samples.	No information available	Concentration and motility analysed by LensHook were comparable to manual assessment.
2019	Hicks et al.	Predict sperm motility from videos and introduce a new dataset.	Sperm motility	A retrospective dataset containing videos of live sperm in untreated samples from 85 subjects (VISEM). Semen analysis was manually evaluated according to WHO 2010.	Deep Learning (CNN)	Deep learning showed potential for rapid and consistent prediction of sperm motility categories (WHO 2010) based on videos of live, untreated sperm samples.

(continued)

Table II Continued

Year	Study	Aim of study	Outcome	Dataset	AI Methods	Summary answer
2019	Riordon *et al.*	Automatic assessment for classification of sperm head morphology into five classes (normal, tapered, pyriform, small, and amorphous).	Sperm morphology	Retrospective images from HuSHeM dataset and 1,132 images from SCIAN dataset.	Deep learning (CNN)	Deep learning can classify sperm head morphology with higher accuracy than previously published AI methods used for the same datasets.
2019	Javadi and Mirroshandel	Automatic assessment of sperm morphology in unfixed cells and introduce a new dataset.	Sperm morphology	1,540 retrospective grey scale images of unfixed sperm cells from 235 subjects (MHSMA dataset). Sperm cells were manually classified as normal or abnormal, and acrosome, head, vacuole, tail, and neck were annotated.	Deep learning (CNN)	The method is able to classify sperm in real-time, but accuracy needs to be improved.
2019	McCallum *et al.*	Automatic method for ranking sperm cells based on DNA quality enabling sperm selection for ICSI.	Sperm DNA integrity	1,064 images of stained sperm cells with known DNA integrity from 6 subjects. It is unclear if the dataset was prospectively collected.	Deep learning (CNN)	Correlation between cell image and DNA integrity was found, and the model was able to predict the DNA integrity of sperm cells in a rapid manner.
2019	Movahed *et al.*	Automatic segmentation of external (head, mid piece, and tail) and internal parts (acrosome and nucleus) of the sperm.	Sperm morphology	A retrospective dataset containing 20 images of stained sperm cells. Sperm parts were manually annotated.	Deep learning (CNN) and traditional ML (Support Vector Machine, K-nearest neighbour, Ensemble Method)	The methods were better at segmenting the head, acrosome, and nucleus than previously described models. Provides the first method for evaluation of tail and mid piece.
2020	Ilhan *et al.*	Fully automated analyses of sperm morphology by a smartphone-based system and introduce a new dataset.	Sperm morphology	200 retrospective images of stained sperm cells from 17 subjects (SMIDS dataset). Sperm cells were manually classified as normal or abnormal.	Deep learning (CNN) and traditional ML (Support Vector Machine, Decision Trees, K-Nearest Neighbours)	The most precise model was able to predict normal or abnormal sperm with an accuracy of 87%.
2021	Abbasi *et al.*	Improve AI models for classification of the sperm head, vacuoles, and acrosome as normal or abnormal.	Sperm morphology	1,540 retrospective images from the MHSMA dataset.	Deep learning (CNN)	Both AI models were able to predict sperm head characteristics more accurately than models previously described in other studies.
2021	Valiuškaite *et al.*	Propose an AI method that can predict if a semen sample is suitable for artificial insemination procedure based on videos of semen samples.	Sperm motility	85 retrospective videos from the VISEM dataset.	Deep learning (CNN)	The AI model detected sperm heads in the videos with an accuracy of 91.8%, and the Pearson correlation between manually assessed motility and predicted sperm head motility was 0.969.

AI, Artificial intelligence; CNN, Convolutional neural network; CASA, Computer-assisted semen analysis.

AI in prediction of outcome before treatment

In several publications, AI was used to build models that predict the possibility of a successful treatment based on a patient's medical record. The result may be of value for patient counselling about the potential results of the treatment. Goyal et al. (2020) used the dataset provided by Human Fertilisation and Embryology Authority (HFEA) which included 30 different features such as age, number of previous ART cycles, number of previous pregnancies, number of inseminated oocytes, number of embryos transferred, and diagnosis for a total of 140 000 patients. Several ML techniques were evaluated to predict live-birth occurrence. They concluded that both male and female traits and living conditions were factors that influenced the outcome of the treatment. A well-known ML technique called extreme gradient boosting (XGBoost) has been used to predict live birth from features such as age, anti-Mullerian hormone, BMI and patient anamnesis (Qiu et al., 2019). Similarly, an ANN was trained to predict live birth using a collection of features such as the age of the female, total dose of gonadotrophins administered, endometrial thickness, and the number of top-quality embryos (Vogiatzi et al., 2019).

AI in analysis of sperm

Most studies using an AI approach for semen analyses have been performed for morphology assessments. The morphological classification is usually performed on stained spermatozoa and implies both distinguishing abnormal from normal spermatozoa as well as identifying various defects of the sperm cell (WHO, 2010). Some of the developed AI models have been trained only to predict the morphology of sperm heads (Chang et al., 2014; Chang et al.; 2017; Shaker et al., 2017; Riordon et al., 2019), whereas other studies describe the recognition of various parts of the whole sperm (Movahed et al., 2019; Ilhan et al., 2020). These differences in the approaches make it difficult to compare results and possible implications for clinical practice even if the overall goal is similar. This is also fortified by the fact that the data used is usually very limited, with only a small number of spermatozoa or patients. Training and evaluating complex methods, for example, DL, with a small-sized dataset most probably leads to an overfitted model. An overfitted model is a model that does not generalise well to unseen real-world cases although it works well on the training data. For example, suppose that a model is trained on a dataset of embryo images to predict pregnancy or not. If the model achieves far higher prediction performance on the embryo images used for training than on new and unseen images, the model is overfitted to the training data.

Annotation of the dataset/sperm images must be done manually and with high accuracy to obtain well-performing models. For recognising and interpreting images of spermatozoa at the pixel level, segmentation is the common approach, in which the spermatozoon is divided into parts, each consisting of a set of pixels. Some studies demonstrate high classification accuracy for morphological characteristics, and most of the studies have both trained and validated the models on freely available datasets, which makes them easier to compare (HuSHeM in Shaker et al. (2017), SCIAN in Chang et al. (2017), and a smaller dataset of 264 spermatozoa in Chang et al. (2014)).

Furthermore, the model performance is compared with existing AI models, and even though this is common practice in the field of AI, it reveals little knowledge about the clinical usability of the model. Regarding sperm morphology, as far as we know, there are no studies comparing the performance of the models with manual assessment according to the WHO guidelines or in relation to fertility outcomes.

For prediction of sperm motility, only one study compared AI-based sperm motility classification against sperm motility that was manually assessed following WHO guidelines (Hicks et al., 2019), while others were mainly focused on comparing various models or exploring the sperm kinematics (Goodson et al., 2017; Valiuškaitė et al., 2020). Studies related to motility and/or morphology also come with the challenge of small datasets, and for both of them, the evaluation procedures are often not clear. Cross validation is sometimes used to compensate for small datasets (Goodson et al., 2017; Shaker et al., 2017). However, even though cross validation is acceptable for testing model performance and comparing it to other models on the same dataset, it does not test the generalisability of the results. In a clinical setting, an independent test set evaluation should be performed, optimally across different clinics (Abbasi et al., 2021).

Automatic systems for diagnostic purposes have been developed. One such system based on an automatic segmentation step and a classification of normal/abnormal spermatozoa has recently been described (Ilhan et al., 2020). The authors reported an accuracy of 87%. However, the method was just compared with other ML methods and not evaluated for its clinical value. In addition, accuracy alone is not a sufficient metric to determine the possible clinical performance of a method, especially if only a small dataset is used. Another automatic system for analysis of sperm concentration, morphology and motility used AI optical microscopic technology, for which the performance was compared with manual assessment (Agarwal et al., 2019, 2021). Nonetheless, the morphology values did not correlate with the manual morphology results, and unfortunately, there are no details provided on the construction and annotation of the dataset.

Parameters that are not part of standard semen analysis have also been used in AI models. For example, sperm intracellular pH was shown to be a stable marker for fertilisation outcome (Gunderson et al., 2021), and sperm DNA integrity could be predicted from bright-field sperm images at a single cell level through supervised training (McCallum et al., 2019). These studies show how AI can be used to automate sperm sorting and selection tasks. However, big datasets from multicentre cohorts are needed to evaluate whether the results are generalisable before these AI models can be used in the clinic as well as for research related purposes. In addition to the conventional semen variables, image features may detect sperm characteristics that are too complex to be recognised by humans, for example, motility patterns or morphological shapes. Nonetheless, from a diagnostic perspective, the clinical value of novel traits must be investigated in epidemiological studies.

The selection of spermatozoa for ICSI is based on a cursory assessment of motility and morphology in real-time, which is especially a challenge for morphology evaluation. The procedure has a potential for improvement using AI to obtain a more objective selection based on the simultaneous monitoring of morphology and motility patterns. Attempts have been made to develop DL models for morphological assessment based on images of unstained spermatozoa (Javadi and Mirroshandel, 2019; Abbasi et al., 2021). Both algorithms can analyse

fresh human sperm in real-time with a magnification between 400× and 600×.

The AI methods used in sperm related studies are mostly based on simple algorithms that are standard implementation in most ML frameworks (Table II). The development of more domain-specific methods and models related to ART will in the long run lead to better results compared to using out-of-the-box methods from existing generic frameworks.

Pitfalls

The AI algorithms are only as good as the data they are based on. There may also be limitations regarding generalisability due to difficulties with the standardisation of the ML methods. Variation in patient demographics, clinical and laboratory practices may cause data bias. When an AI model is based on training in one clinic, the AI model should be validated in independent cohorts (Tran *et al.*, 2019; Bormann *et al.*, 2020b). Furthermore, the models should not be limited to strict inclusion criteria, and optimally the datasets should contain data from different clinics where testing data should be from a different site than the training and validation data (Alegre *et al.*, 2021; Bori *et al.*, 2020).

Another important issue is that patient data and treatment information are not easily obtained for research due to data privacy and ethical considerations. This naturally limits the amount of patient related data to be used for training the AI model. DL methods, which are especially suited for image and video classification, require a large amount of diverse data to be generalisable. Another weakness for some studies is that the data used for training are not connected to any treatment outcome, leading to overly complex models that might only detect irrelevant correlations (Dirvanauskas *et al.*, 2019; Kanakasabapathy *et al.*, 2019; Khosravi *et al.*, 2019; Raudonis *et al.*, 2019; Bormann *et al.*, 2020a; Bormann *et al.*, 2020b; Fukunaga *et al.*, 2020; Rad *et al.*, 2020; Zhao *et al.*, 2021; Alegre *et al.*, 2021). This can raise concerns like, for example, whether the prediction is related to the embryo implantation potential. Moreover, most articles resort to a positive heartbeat at ultrasound control or even a positive hCG test as their outcome, but the most important outcome in ART is the birth of a living, healthy child (Vogiatzi *et al.*, 2019; Bori *et al.*, 2021).

AI models are usually evaluated using different metrics such as accuracy, precision and sensitivity. Often only a small subset or even just a single metric is used to decide if the model performs well. This is not sufficient, and to make a proper estimation about the performance, a set of metrics needs to be considered. It might even be necessary to develop task specific performance measurements.

The future symbiosis between AI and ART

AI methods may be a supporting tool in predicting the patient's individual chance of achieving a healthy child based on available patient data. Adjustments of treatment and prediction of risk and possibilities for complications during pregnancy may be other tasks guided by AI. In ART, AI models may assist in selecting methods, selecting the embryo for transfer, and selecting the spermatozoon for ICSI.

As far as we know, no published studies have performed AI-guided sperm selection for ICSI. Detailed real-time assessment of both motility and morphology simultaneously is a challenge in the present routine. By analysing video recordings of sperm selections by ML methods that consider both the spatial and temporal domains, it may be possible to detect patterns or unknown characteristics that can be related to ICSI outcomes. Similarly, until-now unrecognised features of importance for embryo quality might also be detected by analysing images and videos of embryos.

At present, most of the publications are of a retrospective nature and there is a lack of prospective studies. However, there are some studies that are using retrospective data to perform a prospective study (Bormann *et al.*, 2020b; Huang *et al.*, 2021). The latter should preferably be performed as randomised controlled trials, in which the performance of the AI model included in one arm is compared to decisions routinely performed at a fertility clinic in the other arm, and the outcome is defined as live births. The studies should optimally be designed to include just single embryo transfers to exclude the uncertainty arising when two (or more) embryos are transferred and only one child is born. Most studies using AI for embryo assessment or selection rely on manually extracted features from embryo images or videos. However, over the last couple of years, there has been a rapid increase in the use of DL techniques where features are automatically learned. There are also a few studies using image segmentation techniques to improve automatic embryo assessment (Rad *et al.*, 2020) or to streamline manual assessment (Zhao *et al.*, 2021). The impact of these methods in clinical practice is however limited and standardisation, explainable methods and transparency are keys to improve it.

Standardisation is essential for the development of an applicable and reliable AI model. It requires close interdisciplinary collaboration from the planning of the initial study to the clinical evaluation. In particular, for the successful implementation of AI in the field of ART, a close collaboration between computer science, clinical experience and biological knowledge, which also agree on a common standard, is crucial.

Most algorithms used in all the aforementioned articles, especially DL-based, are black boxes. Ongoing research tries to increase the understanding of these black boxes (Holzinger *et al.*, 2019; Arrieta *et al.*, 2020). In ART, methods for better understanding of black boxes are still in their infancy, focusing on simple visualisation methods (Liu *et al.*, 2020; Abbasi *et al.*, 2021). However, the whole pipeline of an AI system should be transparent (Saito and Rehmsmeier, 2015), including the evaluation method and metrics that need to be described clearly (as in: Javadi and Mirroshandel, 2019; Bori *et al.*, 2020). Increased transparency of AI in ART will also be beneficial for discussions of legal and ethical implications across countries, which often have different regulations.

Furthermore, we need a common way of benchmarking and comparing different systems. In computer science, this is often done using open benchmarking datasets collected and curated by the scientific community. If the hardware changes, like data collected at higher resolutions, the systems will have to be evaluated on the data collected from these new devices. This means we need these community-wide benchmarking datasets to be continuously tested before, during and after clinical trials to verify the performance of AI models. This is not just important for research but also for commercial companies in the

field. Systems such as iDASCORE, KIDScore, Eeva and LensHooke should follow the same requirements and be transparent and open about data, methods and evaluation.

The datasets also need to be continuously updated following technological advances and new findings. There are a few open datasets for sperm and embryo (Shaker et al., 2017; Saeedi et al., 2017; Haugen et al., 2019; Javadi and Mirroshandel, 2019; Ilhan et al., 2020). For sperm, datasets such as VISEM (Haugen et al., 2019) and HuSHeM (Shaker et al., 2017) are commonly used for the evaluation of sperm characteristics. For embryos, even fewer public datasets exist, and the data published by Saeedi et al. (2017) has been used for blastocyst evaluation. Ideally, one publicly available dataset should be used for developing algorithms and a hidden test dataset can be tested on hardware provided by, for example, the European Society of Human Reproduction and Embryology or the American Society for Reproductive Medicine. This would ensure a common standard for training and testing to provide reproducible and comparable results necessary to make AI in ART clinically relevant.

Conclusion

Several studies have applied ML in ART, some of them focusing on clinical relevance, while others concern AI methodological aspects. The limitations are often small datasets and the use of AI algorithms not specifically designed for the fertility clinic. Large open datasets and methods specifically developed and tailored for use in context with ART could lead to better results and understanding.

For AI to significantly impact ART, the model must be developed in the context of clinical practice. Critical steps are proper evaluation and testing of AI systems in relation to outcomes and regulations, a better understanding of the technical aspects, and determination of the performance of AI models regarding practical value in the clinic. In addition, it is important to standardise the use of AI in ART to enable more transparent, comparable, and reproducible results.

To succeed with implementing AI as a valuable tool in the fertility clinic, a strong interdisciplinary collaboration is required between researchers in ART and AI as well as the clinical staff. In addition, there is a need for large-scale randomised controlled trials where several clinics are involved in testing the external validity of the algorithms before defining AI systems that are sufficiently robust for safe clinical implementation.

Data availability

The data generated during and/or analysed during the current study (information extracted from the reviewed articles) are available from the corresponding author on reasonable request.

Authors' roles

M.A.R.: Lead for AI, literature review, writing and revising of text and tables. M.H.S.: Lead for embryo section, literature review, writing and revising. O.W.: Literature search and review, writing and revising. J.M.A.: Tables, figures, literature review, writing. S.A.H.: Tables, figures, literature review, writing and revising of text and tables. H.L.H.: Literature review, writing and revising. E.D.: Literature review, writing. P.H.: Literature review, writing. A.Y.: Literature review, writing. N.H.: Literature review, writing. T.B.H.: Lead for sperm section, literature review, writing and revising.

Funding

The work on this article was partially funded by the Frimedbio project ReproAI granted by the Norwegian Research Council with Project number 288727.

Conflicts of interest

Nothing to disclose.

References

Abbasi A, Miahi E, Mirroshandel SA. Effect of deep transfer and multi-task learning on sperm abnormality detection. *Comput Biol Med* 2021;**128**:104121.

Agarwal A, Henkel R, Huang CC, Lee MS. Automation of human semen analysis using a novel artificial intelligence optical microscopic technology. *Andrologia* 2019;**51**:e13440.

Agarwal A, Panner Selvam MK, Ambar RF. Validation of LensHooke® X1 PRO and computer-assisted semen analyzer compared with laboratory-based manual semen analysis. *World J Mens Health* 2021;**39**:e7.

Alegre L, Del Gallego R, Bori L, Loewke K, Maddah M, Aparicio-Ruiz B, Palma-Govea AP, Marcos J, Meseguer M. Assessment of embryo implantation potential with a cloud-based automatic software. *Reprod Biomed Online* 2021;**42**:66–74.

Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database Syst Rev* 2019; **5**: CD011320. doi: 10.1002/14651858.CD011320.pub4.

Arrieta AB, Diaz-Rodriguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R. *et al.* Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform Fusion* 2020;**58**:82–115.

Bori L, Dominguez F, Fernandez EI, Del Gallego R, Alegre L, Hickman C, Quinonero A, Nogueira MFG, Rocha JC, Meseguer M. An artificial intelligence model based on the proteomic profile of euploid embryos and blastocyst morphology: a preliminary study. *Reprod Biomed Online* 2021;**42**:340–350.

Bori L, Paya E, Alegre L, Viloria TA, Remohi JA, Naranjo V, Meseguer M. Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential. *Fertil Steril* 2020;**114**:1232–1241.

Bormann CL, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, Kandula H, Hariton E, Souter I, Dimitriadis I, Ramirez LB. *et al.* Performance of a deep learning based neural

network in the selection of human blastocysts for implantation. *eLife* 2020a;**9**:1–14.

Bormann CL, Thirumalaraju P, Kanakasabapathy MK, Kandula H, Souter I, Dimitriadis I, Gupta R, Pooniwala R, Shafiee H. Consistency and objectivity of automated embryo assessments using deep neural networks. *Fertil Steril* 2020b;**113**:781–787.

Boulet SL, Mehta A, Kissin DM, Warner L, Kawwass JF, Jamieson DJ. Trends in use of and reproductive outcomes associated with intracytoplasmic sperm injection. *JAMA* 2015;**313**:255–263.

Chang V, Garcia A, Hitschfeld N, Hartel S. Gold-standard for computer-assisted morphological sperm analysis. *Comput Biol Med* 2017;**83**:143–150.

Chang V, Saavedra JM, Castaneda V, Sarabia L, Hitschfeld N, Hartel S. Gold-standard and improved framework for sperm head segmentation. *Comput Methods Programs Biomed* 2014;**117**:225–237.

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* 2020a;**41**: 585–593.

Chavez-Badiola A, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ, Garcia-Sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Sci Rep* 2020b;**10**:4394.

Dang VQ, Vuong LN, Luu TM, Pham TD, Ho TM, Ha AN, Truong BT, Phan AK, Nguyen DP, Pham TN. *et al*. Intracytoplasmic sperm injection versus conventional in-vitro fertilisation in couples with infertility in whom the male partner has normal total sperm count and motility: an open-label, randomised controlled trial. *Lancet* 2021;**397**:1554–1563.

Dirvanauskas D, Maskeliunas R, Raudonis V, Damasevicius R. Embryo development stage prediction algorithm for automated time lapse incubators. *Comput Methods Programs Biomed* 2019; **177**:161–174.

Fukunaga N, Sanami S, Kitasaka H, Tsuzuki Y, Watanabe H, Kida Y, Takeda S, Asada Y. Development of an automated two pronuclei detection system on time-lapse embryo images using deep learning techniques. *Reprod Med Biol* 2020;**19**:286–294.

Goodson SG, White S, Stevans AM, Bhat S, Kao C-Y, Jaworski S, Marlowe TR, Kohlmeier M, McMillan L, Zeisel SH. *et al*. CASAnova: a multiclass support vector machine model for the classification of human sperm motility patterns. *Biol Reprod* 2017; **97**:698–708.

Goyal A, Kuchana M, Ayyagari KPR. Machine learning predicts live-birth occurrence before in-vitro fertilization treatment. *Sci Rep* 2020;**10**:20925.

Gunderson SJ, Puga Molina LC, Spies N, Balestrini PA, Buffone MG, Jungheim ES, Riley J, Santi CM. Machine-learning algorithm incorporating capacitated sperm intracellular pH predicts conventional in vitro fertilization success in normospermic patients. *Fertil Steril* 2021;**115**:930–939.

Haugen TB, Hicks SA, Andersen JM, Witczak O, Hammer HL, Borgli R, Halvorsen P, Riegler M. Visem: a multimodal video dataset of human spermatozoa. In: *Proceedings of the 10th ACM Multimedia Systems Conference*, 2019, 261–266.

Hicks SA, Andersen JM, Witczak O, Thambawita V, Halvorsen P, Hammer HL, Haugen TB, Riegler MA. Machine learning-based analysis of sperm videos and participant data for male fertility prediction. *Sci Rep* 2019;**9**:16770.

Holzinger A, Langs G, Denk H, Zatloukal K, Muller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019;**9**:e1312.

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. *Nat Rev Cancer* 2018;**18**:500–510.

Huang TTF, Kosasa T, Walker B, Arnett C, Huang CTF, Yin C, Harun Y, Ahn HJ, Ohta A. Deep learning neural network analysis of human blastocyst expansion from time-lapse image files. *Reprod Biomed Online* 2021; **42**:1075–1085. doi:10.1016/j.rbmo.2021.02.015.

Høst E, Ernst E, Lindenberg S, Smidt-Jensen S. Morphology of spermatozoa used in IVF and ICSI from oligozoospermic men. *Reprod Biomed Online* 2001;**3**:212–215.

Ilhan HO, Sigirci IO, Serbes G, Aydin N. A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods. *Med Biol Eng Comput* 2020;**58**:1047–1068.

Javadi S, Mirroshandel SA. A novel deep learning method for automatic assessment of human sperm images. *Comput Biol Med* 2019; **109**:182–194.

Kanakasabapathy MK, Thirumalaraju P, Bormann CL, Kandula H, Dimitriadis I, Souter I, Yogesh V, Kota Sai Pavan S, Yarravarapu D, Gupta R. *et al*. Development and evaluation of inexpensive automated deep learning-based imaging systems for embryology. *Lab Chip* 2019;**19**:4139–4145.

Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C. *et al*. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *Npj Digit Med* 2019;**2**:21. doi: 10.1038/s41746-019-0096-y.

Kragh MF, Rimestad J, Berntsen J, Karstoft H. Automatic grading of human blastocysts from time-lapse imaging. *Comput Biol Med* 2019; **115**:103494.

Lemmen JG, Rodriguez NM, Andreasen LD, Loft A, Ziebe S. The total pregnancy potential per oocyte aspiration after assisted reproduction-in how many cycles are biologically competent oocytes available? *J Assist Reprod Genet* 2016;**33**:849–854.

Liu L, Jiao Y, Li X, Ouyang Y, Shi D. Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. *Comput Methods Programs Biomed* 2020;**196**:105624.

McCallum C, Riordon J, Wang Y, Kong T, You JB, Sanner S, Lagunov A, Hannam TG, Jarvi K, Sinton D. Deep learning-based selection of human sperm with high DNA integrity. *Commun Biol* 2019;**2**: 250. doi:10.1038/s42003-019-0491-6.

Milewski R, Kuczyńska A, Stankiewicz B, Kuczyński W. How much information about embryo implantation potential is included in morphokinetic data? A prediction model based on artificial neural networks and principal component analysis. *Adv Med Sci* 2017;**62**:202–206.

Mortimer ST, van der Horst G, Mortimer D. The future of computer-aided sperm analysis. *Asian J Androl* 2015;**17**:545–553.

Movahed RA, Mohammadi E, Orooji M. Automatic segmentation of Sperm's parts in microscopic images of human semen smears using concatenated learning approaches. *Comput Biol Med* 2019;**109**: 242–253.

Paternot G, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and inter-observer analysis in the morphological assessment of early-stage embryos. *Reprod Biol Endocrinol* 2009;**7**:105.doi: 10.1186/1477-7827-7-105.

Qiu J, Li P, Dong M, Xin X, Tan J. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method. *J Transl Med* 2019;**17**:317. doi:10.1186/s12967-019-2062-5.

Rad RM, Saeedi P, Au J, Havelock J. Human Blastocyst's Zona Pellucida segmentation via boosting ensemble of complementary learning. *Inform Med Unlocked* 2018;**13**:112–121.

Rad RM, Saeedi P, Au J, Havelock J. Trophectoderm segmentation in human embryo images via inceptioned U-Net. *Med Image Anal* 2020;**62**:101612.

Raef B, Maleki M, Ferdousi R. Computational prediction of implantation outcome after embryo transfer. *Health Informatics J* 2020;**26**: 1810–1826.

Raudonis V, Paulauskaite-Taraseviciene A, Sutiene K, Jonaitis D. Towards the automation of early-stage human embryo development detection. *BioMed Eng OnLine* 2019;**18**:120. doi: 10.1186/s12938-019-0738-y.

Riordon J, McCallum C, Sinton D. Deep learning for the classification of human sperm. *Comput Biol Med* 2019;**111**:103342.

Saeedi P, Yee D, Au J, Havelock J. Automatic identification of human blastocyst components via texture. *IEEE Trans Biomed Eng* 2017; **64**:2968–2978.

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:e0118432.

Santos Filho E, Noble JA, Poli M, Griffiths T, Emerson G, Wells D. A method for semi-automatic grading of human blastocyst microscope images. *Hum Reprod* 2012;**27**:2641–2648.

Shaker F, Monadjemi SA, Alirezaie J, Naghsh-Nilchi AR. A dictionary learning approach for human sperm heads classification. *Comput Biol Med* 2017;**91**:181–190.

Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: a multicenter study. *Hum Reprod* 2017;**32**:307–314.

Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time-lapse annotations. *Hum Reprod* 2013;**28**:3215–3221.

Tomlinson MJ. Uncertainty of measurement and clinical value of semen analysis: has standardisation through professional guidelines helped or hindered progress? *Andrology* 2016;**4**:763–770.

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**: 1011–1018.

Tsipras D, Santurkar S, Engstrom L, Ilyas A, Madry A. From imagenet to image classification: contextualizing progress on benchmarks. *Int Conference on Machine Learning*, Vol. **119** 2020, 9625–9635.

Valiuškaitė V, Raudonis V, Maskeliūnas R, Damaševičius R, Krilavičius T. Deep learning based evaluation of spermatozoid motility for artificial insemination. *Sensors* 2021;**21**:72.

Vander Borght M, Wyns C. Fertility and infertility: definition and epidemiology. *Clin Biochem* 2018;**62**:2–10.

Ver Milyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, Miller A, Picou A, Murphy AP, Perugini M. 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**:770–784.

Vogiatzi P, Pouliakis A, Siristatidis C. An artificial neural network for the prediction of assisted reproduction outcome. *J Assist Reprod Genet* 2019;**36**:1441–1448.

WHO Laboratory Manual for the Examination and Processing of Human Semen, 5th edn. Genova, Switzerland: WHO Press, 2010. World Health Organization.

Wyns C, Bergh C, Calhaz-Jorge C, De Geyter C, Kupka M, Motrenko T, Rugescu I, Smeenk JA. ART in Europe, 2020: results generated from European registries by ESHRE. *Hum Reprod Open* 2020: hoaa032. doi: 10.1093/hropen/hoaa032.

Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. *World J Gastroenterol* 2019;**25**:1666–1683.

Zhao M, Xu M, Li H, Alqawasmeh O, Chung JPW, Li TC, Lee TL, Tang PMK, Chan DYL. Application of convolutional neural network on early human embryo segmentation during in vitro fertilization. *J Cell Mol Med* 2021;**25**:2633–2644.