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

Running title: Artificial intelligence in the fertility clinic

M. A. Riegler*, Department of Holistic Systems, Simula Metropolitan Center for Digital Engineering, 0167, Oslo, Norway

M. H. Stensen, Fertilitetssenteret, 0167, Oslo, Norway

O. Witczak, Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, 0167, Oslo, Norway

J. M. Andersen, Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, 0167, Oslo, Norway

S. A. Hicks, Department of Holistic Systems, Simula Metropolitan Center for Digital Engineering, 0167, Oslo, Norway

H. L. Hammer, Department of Computer Science, Faculty of Technology, Art and Design, OsloMet - Oslo Metropolitan University, 0167, Oslo, Norway

E. Delbarre, Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, 0167, Oslo, Norway

P. Halvorsen, Department of Holistic Systems, Simula Metropolitan Center for Digital Engineering, 0167, Oslo, Norway

A. Yazidi, Department of Computer Science, Faculty of Technology, Art and Design, OsloMet - Oslo Metropolitan University, 0167, Oslo, Norway

N. Holst, Fertilitetssenteret, 0167, Oslo, Norway

T. B. Haugen, Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, 0167, Oslo, Norway

Abstract

In recent years, the amount of data produced in the field of assisted reproduction technology [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 taking place in medical practice and may become a promising tool to improve the success rate 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.

Keywords

Artificial intelligence, machine learning, assisted reproductive technology, embryology, semen analysis, embryo, spermatozoa, fertility, infertility, algorithm

Introduction

The number of treatments with assisted reproduction technology [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 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 a potential for improvement in the crucial steps in ART treatments, such as selection of embryos for transfer and selection of spermatozoa for intracytoplasmic sperm injection [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 minimize the chance of multiple pregnancies due to the transfer of multiple embryos. Likewise, a more reliable method for sperm selection may increase the success rate 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 for sperm evaluation both for diagnostic purposes and for the decision of the fertilization 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 analyze 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 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 fertilization 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 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).

While 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 a fertilization method (Høst et al., 2001). Strikingly, ICSI is increasingly used irrespective of the male factor infertility diagnosis (Boulet et al., 2015; Vander Borght et al., 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 ofhe 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 contribute with 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 et al., 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 remains debated (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 prone to error while it is 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 analyzing 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 Figure 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 Figure 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 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 unlabeled data, for example, automatically grouping blastocyst images based on visual features that may correlate with morphological characteristics. 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 (Kanakasabapathy et al., 2019; Khosravi et al., 2019; Raudonis et al., 2019; Dirvanauskas et al., 2019; Fukunaga et al., 2020; Bormann et al., 2020a; Bormann et al., 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 utilize 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, 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 generalization 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. Furthermore, when evaluating implantation rates for embryos grouped by morphology grades, a CNN provided a slightly higher correlation between predicted embryo quality and implantability than human embryologists (Kragh et al., 2019). The use of a CNN trained to assess an embryo's implantation potential directly by 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 doubleblinded 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 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' accuracy and 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 multicentre nature of the study supported its applicability at different clinics, standardizing the interpretation of embryo development.

Embryo assessment, ranking, and selection are procedures normally based on evaluations at different time points in the 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; Bormann et al., 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 labeling 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; Tran et al., 2019; Dirvanauskas 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).

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 counseling 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 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, the 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 spermatozoon from normal as well as identifying various defects of the 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 a model that memorized the data (overfitting) without being generalizable.

Annotation of the dataset/sperm images must be done manually and with high accuracy to obtain well-performing models. For recognizing 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., 2021). 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 generalizability 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; Agarwal et al., 2021). Nonetheless, the morphology values did not correlate with 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 fertilization outcome (Gunderson et al., 2019), and sperm DNA integrity could be predicted from brightfield 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 if the results are generalizable before these AI models can be used in the clinic, and not only for research related purposes. In addition to the conventional semen variables, image features may detect sperm characteristics that are too complex to be recognized 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 by the use of AI to obtain a more objective selection based on the simultaneous monitoring of morphology and motility pattern. 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 are able to analyze fresh human sperm in real-time with a magnification between 400x and 600x.

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 generalizability due to difficulties with the standardization 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 use 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 generalizable. Another weakness for some studies is that the data used for training is not connected to any treatment outcome, leading to overly complex models that might only detect irrelevant correlations (Kanakasabapathy et al., 2019; Khosravi et al., 2019; Raudonis et al., 2019; Dirvanauskas et al., 2019; Bormann et al., 2020a; Bormann et al., 2020b; Rad et al., 2020; Fukunaga 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 positive heartbeat at ultrasound control or even positive urine 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 taken into account. 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 analyzing 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 unrecognized features of importance for embryo quality might also be detected by analyzing images and videos of embryos.

At present, most of the publications are of retrospective nature and there is a lack of prospective studies. The latter should preferably be performed as randomized 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. 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 standardization, explainable methods, and transparency are keys to improve it.

Standardization 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 (Arrieta et al., 2020; Holzinger et al., 2019). In ART, methods for better understanding of black boxes are still in their infantile, focusing on simple visualization methods (Abbasi et al., 2021; Liu et al., 2020). 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 like 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.

The datasets also need to be continuously updated following technological advances and new findings. There are a few open datasets for sperm and embryo (Haugen et al., 2019; Shaker et al., 2017; Saeedi et al., 2017; Ilhan et al., 2020; Javadi and Mirroshandel, 2019). 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, where the data published by Saeedi et al. (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. 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 tailored for the 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 how to determine the performance of AI models of practical value in the clinic. In addition, it is important to standardize 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 largescale randomized 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.

Author's roles

Michael A. Riegler: Lead AI, literature review, writing

Mette Haug Stensen: Lead embryo, literature review, writing

Oliwia Witczak: Literature search and review, writing

Jorunn M. Andersen: Tables, figures, literature review, writing

Steven A. Hicks: Tables, figures, literature review, writing

Hugo L. Hammer: Literature review, writing

Erwan Delbarre: Literature review, writing

Pål Halvorsen: Literature review, writing

Anis Yazidi: Literature review, writing

Nicolai Holst: Literature review, writing

Trine B. Haugen: Lead sperm, literature review, writing

Data Availability Statement

No new data were generated or analyzed in support of this research.

Conflicts of interest

Nothing to disclose

Acknowledgements

Funding

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

Table Headings:

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

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

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;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. 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 *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;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 on a 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;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;doi:10.1038/s42003-019-0491-6.

Milewski R, Kuczynska A, Stankiewicz B, Kuczynski 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. *Advances in Medical Sciences* 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;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;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;doi:10.1186/s12938-019-0738-y.

Riordon J, McCallum C, Sinton D. Deep learning for the classification of human sperm. *Comput Biol and 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. *International Conference on Machine Learning*. 2020;119:9625-9635.

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* 2021;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.

World Health Organization. *WHO laboratory manual for the examination and processing of human semen*. 5th edn. Genova, Switzerland. WHO Press. 2010.

Wyns C, Bergh C, Calhaz-Jorge C, De Geyter C, Kupka M, Motrenko T, Rugescu I, Smeenk J, Tandler-Schneider A. ART in Europe, 2016: results generated from European registries by ESHRE. *Hum Reprod Open* 2020;doi: https://doi.org/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.

EVERYTHING BELOW IS BROKEN DONT TOUCH WILL DESYNCH DOCUMENT