Non-invasive prediction of human embryonic ploidy using artificial intelligence: a systematic review and meta-analysis

Xing Xin,[a,b,f] Shanshan Wu,[a,b,f] Heli Xu,[c,f] Yujiu Ma,[a,b,f] Nan Bao,[d] Man Gao,[e] Xue Han,[e] Shan Gao,[a,b] Siwen Zhang,[a,b] Xinyang Zhao,[a,b] Jiarui Qi,[a,b] Xudong Zhang,[a,b] and Jichun Tan[a,b,*]

[a]Centre of Reproductive Medicine, Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, No. 39 Huaxiang Road, Tiexi District, Shenyang 110022, China
[b]Key Laboratory of Reproductive Dysfunction Disease and Fertility Remodeling of Liaoning Province, No. 39 Huaxiang Road, Tiexi District, Shenyang 110022, China
[c]Department of Clinical Epidemiology, Shengjing Hospital of China Medical University, Shenyang 110022, China
[d]The College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110167, China
[e]Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, No. 36 Sanhao Street, Heping District, Shenyang 110004, China

eClinicalMedicine
2024;77: 102897
Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102897

Summary

Background Embryonic ploidy is critical for the success of embryo transfer. Currently, preimplantation genetic testing for aneuploidy (PGT-A) is the gold standard for detecting ploidy abnormalities. However, PGT-A has several inherent limitations, including invasive biopsy, high economic burden, and ethical constraints. This paper provides the first comprehensive systematic review and meta-analysis of the performance of artificial intelligence (AI) algorithms using embryonic images for non-invasive prediction of embryonic ploidy.

Methods Comprehensive searches of studies that developed or utilized AI algorithms to predict embryonic ploidy from embryonic imaging, published up until August 10, 2024, across PubMed, MEDLINE, Embase, IEEE, SCOPUS, Web of Science, and the Cochrane Central Register of Controlled Trials were performed. Studies with prospective or retrospective designs were included without language restrictions. The summary receiver operating characteristic curve, along with pooled sensitivity and specificity, was estimated using a bivariate random-effects model. The risk of bias and study quality were evaluated using the QUADAS-AI tool. Heterogeneity was quantified using the inconsistency index (I^2), derived from Cochran's Q test. Predefined subgroup analyses and bivariate meta-regression were conducted to explore potential sources of heterogeneity. This study was registered with PROSPERO (CRD42024500409).

Findings Twenty eligible studies were identified, with twelve studies included in the meta-analysis. The pooled sensitivity, specificity, and area under the curve of AI for predicting embryonic euploidy were 0.71 (95% CI: 0.59–0.81), 0.75 (95% CI: 0.69–0.80), and 0.80 (95% CI: 0.76–0.83), respectively, based on a total of 6879 embryos (3110 euploid and 3769 aneuploid). Meta-regression and subgroup analyses identified the type of AI-driven decision support system, external validation, risk of bias, and year of publication as the primary contributors to the observed heterogeneity. There was no evidence of publication bias.

Interpretation Our findings indicate that AI algorithms exhibit promising performance in predicting embryonic euploidy based on embryonic imaging. Although the current AI models developed cannot entirely replace invasive methods for determining embryo ploidy, AI demonstrates promise as an auxiliary decision-making tool for embryo selection, particularly for individuals who are unable to undergo PGT-A. To enhance the quality of future research, it is essential to overcome the specific challenges and limitations associated with AI studies in reproductive medicine.

Funding This work was supported by the National Key R&D Program of China (2022YFC2702905), the Shengjing Freelance Researcher Plan of Shengjing Hospital and the 345 talent project of Shengjing Hospital.

Keywords: Artificial intelligence; Deep learning; Machine learning; Embryonic ploidy

*Corresponding author. Centre of Reproductive Medicine, Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University, No. 39 Huaxiang Road, Tiexi District, Shenyang, 110022, China.
E-mail address: tjczjh@163.com (J. Tan).
[f]These authors contributed equally.

Research in context

Evidence before this study

Embryonic ploidy is crucial for successful embryo transfer, yet current non-invasive methods for predicting ploidy remain limited in accuracy. Advances in artificial intelligence (AI) offer a promising solution to improve predictive performance. However, no quantitative synthesis has yet comprehensively assessed the effectiveness of AI in predicting embryo ploidy. To fill this gap, we conducted a systematic review and meta-analysis, performing a comprehensive literature search across multiple databases up to August 10, 2024, without language restrictions.

Added value of this study

To our knowledge, this is the first systematic review and meta-analysis focused on AI-based prediction of embryonic ploidy from imaging data. We adhered strictly to diagnostic review guidelines and conducted a comprehensive search across medical and engineering databases to ensure thorough coverage. Our findings suggest that AI shows strong potential as a decision-making tool for embryo selection, particularly for patients unable to undergo preimplantation genetic testing for aneuploidy (PGT-A).

Implications of all the available evidence

AI algorithms show promising performance in predicting embryonic ploidy from imaging data. While current models cannot fully replace invasive methods for determining ploidy, AI offers the potential as a valuable decision-making tool for embryo selection, especially for individuals unable to undergo PGT-A. Adopting more rigorous reporting standards that address the unique challenges inherent in AI research would be instrumental in enhancing the quality and reliability of future studies.

Introduction

Embryo aneuploidy is a primary contributor to embryonic dysplasia, implantation failures, pregnancy losses, and congenital abnormalities in newborns.[1,2] During *in vitro* fertilization (IVF) treatments, the aneuploidy rates in two-pronuclear stage embryos typically range from 25% to 40%, escalating with maternal age.[3,4] Preimplantation genetic testing for aneuploidy (PGT-A) employs biopsy techniques for precise chromosomal assessment,[5,6] enabling embryologists to ascertain embryo ploidy before transfer, thus improving pregnancy outcomes of IVF treatment. Nevertheless, several limitations hinder the practical application of PGT-A. Embryo biopsy, an invasive procedure, may damage embryos and reduce their developmental potential.[7] Legal and ethical restrictions further restrict access to PGT-A for some patients.[8] Additionally, not all embryos are suitable for biopsy, limiting the applicability of this technology. Economic factors also pose significant constraints; for example, PGT-A costs can exceed £3000 in the UK and $12,000 in the US, affecting its accessibility and adoption.[9] Consequently, research is increasingly focusing on non-invasive techniques for embryo ploidy testing, aiming to provide viable alternatives to PGT-A.

Time-lapse systems, which could capture detailed multiplanar images of the embryonic development process at regular intervals, are increasingly utilized globally in the field of reproductive medicine, primarily for embryo quality assessment during IVF treatments.[10] Previous studies have demonstrated significant correlations between morphokinetic variables and embryo euploidy.[11,12] The integration of time-lapse videography into IVF could provide detailed annotations of embryo morphokinetics and facilitate the identification of novel biomarkers for embryo selection. However, relying solely on morphokinetic parameters to predict embryo euploidy still presents significant challenges due to considerable variability between aneuploid and euploid embryos.[13]

Advancements in artificial intelligence (AI) could potentially bridge the significant gap between the high demand for non-invasive predictions of embryo ploidy and the currently limited predictive accuracy of such assessments.[14,15] Radiomics, a novel data-driven approach, extracts numerous quantitative features from medical images.[16] These features can be analysed using machine learning (ML) or deep learning (DL) techniques.[17] ML, a subset of AI, minimizes operator subjectivity and enhances the accuracy of embryonic ploidy prediction. Yuan et al. developed a robust ML model that integrated the morphokinetic and morphological characteristics of blastocysts with patients' clinical parameters to predict the euploidy of blastocysts and the area under the curve (AUC) of this model reached 0.879, indicating its high predictive accuracy.[18] In recent years, the clinical application of DL in embryonic ploidy prediction has surged.[19,20] An AI model constructed by S.M. Diakiw et al. using a Convolutional Neural Network (CNN) achieved an accuracy of 77.4% in predicting embryonic euploidy.[19] However, DL, characterized by artificial neural networks with multiple hidden layers, often lacks transparency due to its complex structure, leading to ethical and societal concerns among IVF professionals due to its 'black-box' nature.[21,22] Moreover, researchers are constantly exploring various methods to enhance diagnostic accuracy, including improving image quality, incorporating more clinical data of patients, increasing sample sizes, and optimizing algorithms.[23]

The research on AI in reproductive medicine employs a diverse range of methodologies, from DL frameworks analysing morphokinetic variables to advanced algorithms that integrate clinical and imaging data. Despite promising advancements in AI, the performance inconsistency of these models, the variability in study designs, and the constraints posed by limited dataset sizes are notable challenges. To address these issues, a comprehensive systematic review is imperative to thoroughly assess the effectiveness, dependability, and feasibility of AI applications in embryo selection. This study leverages meta-analysis to bridge existing research gaps, enhancing our understanding of the strengths and limitations of non-invasive, image-based ploidy prediction techniques. The findings could revolutionize clinical practices by offering a less invasive and auxiliary decision-making tool for embryo selection, thereby enhancing the safety, efficiency, and accessibility of ART for a wider patient demographic.

Methods

Protocol registration and study design

The study was officially registered with PROSPERO(CRD42024500409). The meta-analysis adhered to established reporting standards, specifically the PRISMA[24] and CHARMS [25] reporting guidelines.

Search strategy and eligibility criteria

A comprehensive literature search was performed using the following databases: PubMed, MEDLINE, Embase, IEEE, SCOPUS, Web of Science, and the Cochrane Central Register of Controlled Trials. These databases represent the entirety of our search scope, ensuring broad coverage across medical, engineering, and technology fields. This systematic review targeted studies that developed AI algorithms to evaluate the diagnostic performance of human embryonic ploidy using medical imaging techniques. The literature search was limited to articles published up to August 10, 2024, without language restrictions. The search strategy employed across all databases included the following terms: ('Artificial intelligence' OR 'Machine learning' OR 'Deep learning' OR 'Neural network') AND ('Performance' OR 'Sensitivity' OR 'Specificity' OR 'Accuracy' OR 'Area under the curve') AND ('Genetic testing' OR 'Genetic screening' OR 'Preimplantation genetic testing' OR 'Preimplantation genetic screening' OR 'Preimplantation genetic diagnosis' OR 'Embryo') AND ('Chromosomal constitution' OR 'Aneuploid*' OR 'Euploid*' OR '*ploid*'). The asterisk (*) serves as a wildcard, allowing the search engine to include any relevant auto-completion of the specified search term. A detailed summary of the search strategies employed for each database is provided in Supplementary Note 1.

In this systematic review, we considered studies evaluating the efficacy of AI models in the non-invasive prediction of human embryonic ploidy. Eligible studies reported on any outcomes such as accuracy, sensitivity (Se), specificity (Sp), positive predictive value, and negative predictive value, or provided detailed data from 2 × 2 contingency tables. We imposed no restrictions concerning participant characteristics or the specific contexts in which AI models were applied. Both prospective and retrospective research designs were included. Exclusion criteria were as follows: (1) duplicate publications; (2) letters to the editor, editorials, conference abstracts, systematic reviews or meta-analyses, consensus statements, guidelines, and review articles; (3) studies not pertinent to the designated topic; (4) studies utilizing non-human samples; (5) studies lacking an AI model. Two reviewers (XX and Y-JM) independently screened titles and abstracts based on these criteria. Full texts of potentially relevant articles were subsequently retrieved for detailed evaluation. Any disagreements were discussed with a third reviewer (S-SW) and resolved through consensus.

Data extraction

In the systematic review process, data regarding study characteristics and diagnostic performance were independently extracted by two reviewers (XX and Y-JM) utilizing a standardized data extraction form (Tables 1–4), which was carefully developed to ensure comprehensive and accurate data collection. The form included key variables relevant to our study objectives and was structured to capture all necessary information systematically. To address consistency between reviewers, we conducted a pilot test of the form. Any discrepancies that arose during this phase were addressed through discussion between the two primary reviewers. In instances where consensus could not be achieved, a third investigator (S-SW) was consulted to resolve the disagreements.

In the systematic review, diagnostic accuracy metrics —true positive (TP), false positive (FP), false negative (FN), and true negative (TN)—were collated directly into contingency tables. These tables facilitated the computation of Se and Sp. In instances where a single study provided multiple contingency tables corresponding to the same or different AI algorithms, each was treated as an independent observation.[43,44] Supplementary Table S1 provides a summary of the contingency tables derived from the included studies. For studies where contingency table data were not available from the original publication, we contacted the authors via email to request the raw data. Ultimately, eight studies did not yield the necessary data and were therefore excluded from the meta-analysis.

Study quality assessment

In the process of quality assessment, each study selected for inclusion underwent evaluation using the quality assessment of diagnostic accuracy studies for artificial

Author, year	Inclusion criteria	Exclusion criteria	Number of embryos	Number of patients	Mean or median age (SD; range)	COH protocol	Algorithm architecture
T. Bamford et al.,[26] 2023[a]	Patients selected for PGT-A primarily for advanced maternal age, recurrent implantation failure (>2 failed embryo transfers), recurrent miscarriage (>2 spontaneous miscarriages), or to shorten the time to pregnancy.	Mosaic embryos were excluded from the initial modelling.	Dataset 1 = 8027 embryos (3004 euploid and 5023 aneuploid); Dataset 2 = 2457 embryos (1008 euploid and 1449 aneuploid)	1725	NR	long GnRH agonist or short antagonist protocol.	ML (LR, RFC, XGBoost)/DL (DL)
J. Barnes et al.,[27] 2023[a]	Embryos that were biopsied for PGT-A either on day 5 if the morphological grade was 2BB or better, or by day 6 if they reached the blastocyst stage, using the Veeck and Zaninovic grading system. In instances where patients had a small number of viable embryos, embryos were biopsied on day 6 even if they were in the morula stage or the cavitating morula stage.	Embryos with significant missing morphokinetic parameters. Underexposed static images were removed from the dataset. Mosaic embryos were excluded from the final dataset.	10,378 embryos (4425 euploid and 5953 aneuploid)	1385	36.98 (4.62)	NR	ML (XGBoost, k-NN, SVM, RF)/DL (ResNet18 CNN)
A. Chavez-Badiola et al.,[28] 2020[a]	More than one blastocyst available; PGT-A test results available; and having both euploid and aneuploid blastocysts within the same cycle.The micrographs passed through a series of quality filters, including sufficient light for clear visibility, sharp focus of the zona pellucida and trophectoderm, one embryo per micrograph, the entire embryo shown within the limits of the image.	With visible instruments or debris, hindrance of visibility by text or symbols in the images.	840 embryos	NR	37.1 (36.5–37.7)	NR	DL (deep neural network)
B. Huang et al.,[29] 2021[a]	All embryo images are captured during 5 or 6 days after fertilization before biopsy by time-lapse microscope system.	Cycles where the patient gave up PGT were excluded.	1490 blastocysts (617 euploid, 873 aneuploid)	469	30.8 (4.5)	NR	DL (3D CNN)
C. I. Lee et al.,[30] 2021[a]	Patients undergoing PGT-A were selected. The dataset comprised time-lapse videos with known outcomes from PGT-A, capturing embryo development from day 1 to day 5.	AMH $\leq$ 1.1 ng/mL, advanced age group (>38 years old), severe endometriosis and uterine pathology, surgical sperm retrieval, and the patient experienced at least three previous failures of euploid embryo transfers.	690 blastocysts (533 euploid and mosaicism, 157 aneuploid)	108	NR	NR	DL (Two-Stream Inflated 3D ConvNet)
S. De Gheselle et al.,[31] 2022[a]	The inclusion criteria for embryos in this study involved undergoing trophectoderm biopsy for PGT, being part of an ICSI cycle, and having detailed embryonic development and morphokinetic data recorded through time-lapse imaging.	NR	539 blastocysts (244 euploid, 295 aneuploid)	128	34.0 (22.0–43.0)	Ovarian stimulation was performed with either GnRH agonist (short- or long-acting) or antagonist protocol.	ML (RCF, GB, SVM, NB, MLR)
S. M. Diakiw et al.,[19] 2022[a]	Female patients aged at least 18 years who underwent IVF procedures. Included images of embryos taken using optical light microscopy systems, with matched PGT-A results as the ground truth outcome. All images were required to have a minimum resolution of 480 × 480 pixels with the complete embryo in the field of view and the focus on the inner cell mass (ICM).	PGT-A result was inconclusive or missing. Technical issues (duplicate images, unmatched images/ metadata, etc.). Day 6 embryos (excluded from training but used to evaluate model performance, as described). Days other than Day 5 or Day 6 (or day not recorded). Mosaic embryos (excluded from training but used to evaluate model performance, as described).	5050 embryos (3251 euploid, 1799 aneuploid)	2438	36.2 (19–53)	NR	DL (CNN: DenseNet-161, ResNet-50, DenseNet-121)

(Table 1 continues on next page)

Author, year	Inclusion criteria	Exclusion criteria	Number of embryos	Number of patients	Mean or median age (SD; range)	COH protocol	Algorithm architecture
(Continued from previous page)							
Y. Zou et al.,[32] 2022[a]	Only the embryos graded over 5BC or 5CB, according to criteria previously described by Gardner and Lane, were defined as available blastocysts and were taken into chromosomal analysis.	Embryos that were not suitable for time-lapse assessment, because excessive cytoplasmic fragmentations at the cleavage stage (>50%) produced a poor-quality time-lapse image, were excluded.	773 embryos (358 euploid, 415 aneuploid)	212	22–46	Ovarian stimulation of the patients was induced by short gonadotrophin releasing hormone agonist protocol (43.2%) or gonadotrophin releasing hormone antagonist protocol (56.8%).	ML (DT, RF, GBDT, Adaboost, XGBoost)/DL (DNN, DNN–LSTM)
G. B. Danardono et al.,[33] 2023[a]	The embryos had their ploidy status determined through PGT-A, and only those images where the embryos were fully captured without additional objects like holding or biopsy pipettes and were clear were included in the study.	Images with obstructions, poor quality images.	865 blastocysts (196 euploid, 331 aneuploid, 416 mosaic)	483	35 (32–39)	NR	ML (DT, RF, GB, SVM, LR) and DL (CNN)
E. Paya et al.,[20] 2023[a]	Embryos subjected to PGT-A were cultured using the EmbryoScope or EmbryoScope Plus TL systems.	Mosaic embryos were excluded from the study.	1151 blastocysts (493 euploid, 658 aneuploid)	NR	NR	ovarian stimulation was performed with GnRH agonist treatment	DL (CNN, BiGRU, BiLSTM, GRU, LST, ResNet50)
Z. Yuan et al.,[18] 2023[a]	The patients included in this study had to meet specific clinical indications, such as recurrent implantation failure, recurrent pregnancy loss, severe teratozoospermia, or female advanced age, defined as being 38 years or older and requiring assisted reproductive technology. All patients were required to have normal chromosomal karyotypes.	NR	1396 blastocys (877euploid, 519 aneuploid)	403	35.47 (4.86)	Long-acting or antagonist protocols.	ML (LR)
T. M. Luong et al.,[34] 2024[a]	The inclusion criteria for the study were embryos from IVF cycles that underwent PGT-A and used time-lapse incubation.	Mosaic embryos andembryos with insufficient DNA quality were excluded from the analysis	1908 embryos (692 euploid, 1216 aneuploid)	820	38.5 (3.85)	GnRH antagonist, long GnRH agonist, ultrashort GnRH agonist, PPOS, or mild stimulation.	ML (RF, LDA, LR, SVM, AdaBoost, LGBM)
J. A. Ortiz et al.,[35] 2023	Indications for PGT-A included advanced maternal age, abnormal karyotype of one of the parents, high rate of chromosome aneuploidies in sperm samples, history of chromosomal abnormalities, repeated miscarriages, and embryo implantation failure. Embryos included in the study were from IVF cycles that had undergone PGT-A on day-5, -6, or -7.	Noninformative embryos were excluded from the analysis.	6989 blastocysts (3731 euploid, 3258 aneuploid)	2476	33.82 (6.82)	NR	ML (RF)
F. Chen et al.,[36] 2023	Couples that were part of PGT cycles, specifically those involving PGT-SR. Blastocysts included were those that received PGT treatment and had comprehensive chromosome screening (CCS) results available.	Embryos or cycles lacking complete time-lapse videos or comprehensive chromosome screening results were excluded. Embryos with severe gene disorders were not included in the PGT-A or PGT-SR testing, hence were excluded from the study.	1422 blastocysts (651 euploid, 771 aneuploid)	355	31.2 (4.6)	COS was accomplished by GnRH agonist suppression protocol, GnRH antagonist flexible protocol, or micro stimulation protocol.	DL (RN)
V. S. Jiang et al.,[37] 2023	NR	Embryos with completely non-discernable images were removed from the study.	699 blastocysts (339 euploid, 360 aneuploid)	248	37.30 (3.6)	NR	DL (CNN)

(Table 1 continues on next page)

Author, year	Inclusion criteria	Exclusion criteria	Number of embryos	Number of patients	Mean or median age (SD; range)	COH protocol	Algorithm architecture
(Continued from previous page)							
S. Rajendran et al.,[38] 2023	The study included embryos with available time-lapse sequences of development and relevant clinical data such as embryologist-derived blastocyst scores, morphokinetic parameters, and maternal age at the time of oocyte retrieval.	Sequences were excluded if the embryo was absent from the petri dish, less than half-visible, or the image was too dim to discern the embryo.	1998 blastocysts (916 euploid, 1082 aneuploid)	498	NR	NR	DL (VGG16 CNN, BiLSTM)
L. Sun et al.,[39] 2024	Embryos with available static images and time-lapse videos. Patients involved in PGT cycles.	Embryos that did not meet specific morphological criteria (such as blastocysts that were not stage ≥3, or inner cell mass or trophectoderm score < B)	145 embryos	543	NR	NR	ML (RF) and DL (CNN)
N. Handayani et al.,[40] 2024	Recurrent IVF failure following the transfer of top-quality embryo(s), having a history of recurrent miscarriages, and advanced maternal age.	PGT-A sample failed to pass quality control, required re-biopsy, or yielded undetermined results.	1020 blastocysts (181 euploid, 390 aneuploid, 449 mosaic)	425	36 (7), median (IQR)	NR	DL (CNN)
B. X. Ma et al.,[41] 2024	NR	Any samples that failed DNA amplification during the NGS analysis were excluded from the study.	3405 blastocysts (1464 euploid, 1522 aneuploid, 419 mosaic)	979	31.78 (4.44)	NR	DL (iDAScore)
H. He et al.,[42] 2024	Patients with ICSI-fertilized embryos, embryos with continuous cleavage-stage culture, and embryos with blastocyst culture during fresh cycles were the only ones included	Patients with systemic immune illnesses such as thyroiditis, systemic lupus erythematosus, aberrant morphological oocytes, scarred uteri, uterine deformity, uterine adhesions, and other organic uterine problems were also disqualified.	184 blastocysts (95 euploid, 89 aneuploid)	NR	NR	NR	ML (LR, LGBM, XGBoost, CatBoost, RF)

AdaBoost: adaptive boosting; BiGRU: Bidirectional gate recurrent unit; BiLSTM: Bidirectional long short-term memory; CNN: convolutional neural network; COH: controlled ovarian hyperstimulation; DL: deep learning; DNN: deep neural networks; DT: decision tree; GB: gradient boosting; GBDT: gradient boosting decision tree; GRU: gate recurrent unit; GnRH: gonadotropin-releasing hormone; iDAScore: intelligent data analysis score; k-NN: k-nearest neighbour; LDA: Linear discriminant analysis; LGBM: Light gradient-boosting machine; LR: Logistic regression; LSTM: long short-term memory; MLR: multivariable logistic regression; NB: naïve Bayes (gaussian); NR: not reported; PGT-A: preimplantation genetic testing for aneuploidy; PGT-SR: preimplantation genetic testing for chromosomal structural rearrangements; RFC: random forest classifier; RF: Random Forest; RN: ResNet50; SVM: support vector machine; XGBoost: extreme gradient boosting. [a]Studies (n = 12) included in the meta-analysis.

Table 1: **Participant demographics and algorithm architecture for the 20 included studies.**

intelligence (QUADAS-AI) criteria,[45] conducted independently by two reviewers (XX and MG). Detailed outcomes of these assessments are available in Supplementary Table S2. The QUADAS-AI tool is designed to equip researchers with a tailored framework for assessing the risk of bias and applicability in reviews focused on AI diagnostic test accuracy, as well as in comparative accuracy studies that include at least one AI-based index test. Any discrepancies between reviewers were resolved through consultation with a third collaborator (XH).

Statistics

In the study, the primary outcomes were Se, Sp, and AUC. The hierarchical summary receiver-operating characteristic (SROC) curve was utilized to ascertain the precision of the AI model. The SROC curve, inclusive of the corresponding 95% confidence region and 95% prediction region, was constructed around the averaged Se, Sp, and AUC estimates. When multiple AI models were evaluated within a single study, the model demonstrating the highest accuracy was selected for subsequent meta-analytic procedures.

Spearman correlation test for the presence of diagnostic threshold effect. Given the anticipated diversity across studies, a bivariate random effects model was applied with both sensitivity and specificity were transformed using the logit transformation before performing the meta-analysis.[46] The forest plot illustrates the heterogeneity across the included studies. Substantial heterogeneity is indicated by an inconsistency index (I^2) ≥ 50%, or a p-value of ≤0.10 based on Cochran's Q test.[47,48] The relationship between Se and Sp was further examined through a bivariate boxplot.[49] A sequential sensitivity analysis was conducted by sequentially excluding individual studies to assess the robustness of the findings and evaluate their impact on heterogeneity and diagnostic performance metrics.[50] To identify potential sources of heterogeneity, meta regression analyses were undertaken. The predictors assessed in this study included algorithm type, AI-driven Decision Support Systems (DSS), annotation methods, external

Author, year	Gold standard (genetic platform)	Detection thresholds for euploidy, aneuploidy and mosaic	Type of internal validation	External validation	Number of embryos for training/ validation/testing
T. Bamford et al.,[26] 2023[a]	PGT-A (aCGH in 367 (5%) and NGS in 7660 (95%) blastocysts)	Samples with <20% aneuploidy were classed as euploid, 20–80% mosaic, and >80% aneuploid.	Internal–external cross-validation	Yes	Dataset 1: 6420/NR/ 1607 (8:2); Dataset 2: 1967/NR/490 (8:2)
J. Barnes et al.,[27] 2023[a]	PGT-A	NR	Five-fold cross-validation	Yes	8336/1557/1555
A. Chavez-Badiola et al.,[28] 2020[a]	PGT-A	NR	Split-sample validation	No	680/76/84 (8:1:1)
B. Huang et al.,[29] 2021[a]	PGT-A (NGS: Life Technologies Ion Proton)	The threshold for aneuploidy detection was set to be greater than 70%. The threshold for mosaic detection varies from chromosomes. For chromosomes 13, 16, 18, and 21, the lower limit was 30%, for the 19 chromosome, lower limit was 50%, for others, lower limit was 40%. The value below the lower limit indicates a euploidy.	Ten-fold cross-validation	Yes	921/102/467
C. I. Lee et al.,[30] 2021[a]	PGT-A (hr-NGS)	Euploid blastocysts with mosaicism levels ≤20%; low-level mosaic blastocysts with mosaicism levels between 20 and 50%; high level mosaic blastocysts with mosaicism levels between 50 and 80%; aneuploid blastocysts with mosaicism levels >80%.	NR	No	552/NR/138 (8:2)
S. De Gheselle et al.,[31] 2022[a]	PGT-A or PGT-SR (NGS)	NR	Ten-fold cross-validation	No	388/43/108
S. M. Diakiw et al.,[19] 2022[a]	PGT-A	NR	Split-sample validation	YES	3174/300/1001 (7:1:2)
Y. Zou et al.,[32] 2022[a]	PGT (SNP array testing.)	Embryos less than 20% of the mosaic were considered euploid, and those more than 80% of the mosaic were considered aneuploid (Xiao et al., 2021). The other embryos (20–80% aneuploid) were classified as mosaic and were excluded in the study.	Five-fold cross-validation	No	494/124/155
G. B. Danardono et al.,[33] 2023[a]	PGT-A (NGS)	The threshold for calling mosaic was a 30%–80% mixture of euploid and aneuploid cells (<30% was euploid, and >80% was aneuploid).	Split-sample validation	No	692/NR/173 (8:2)
E. Paya et al.,[20] 2023[a]	PGT-A (NGS).	NR	Split-sample validation	No	932/104/115 (8:1:1)
Z. Yuan et al.,[18] 2023[a]	PGT-A	NR	NR	No	NR/NR/1396
T. M. Luong et al.,[34] 2024[a]	PGT-A (NGS)	Embryos with less than 30% full chromosome aneuploidy in the biopsy were classified as euploid, while those with more than 70% full chromosome aneuploidy were categorized as aneuploid. Embryos displaying 30–70% of the presence of two or more cell lines with chromosomal complements were classified as mosaics and were excluded from the study.	Five-fold cross-validation	Yes	1176/437/295
J. A. Ortiz et al.,[35] 2023	PGT-A (aCGH, NGS)	Embryos with a percentage of aneuploid cell line <25% were classified as euploid. Embryos were classified as mosaic if the percentage of the aneuploid cell line was between 25% and 50%. , if the proportion of aneuploid cells was >50%, the embryo was classified as aneuploid.	Ten-fold cross-validation	No	6290/699/1398
F. Chen et al.,[36] 2023	PGT-A or PGT-SR (SNP microarray, NGS)	Euploid blastocysts with mosaicism levels <50%; aneuploid blastocysts including numerical chromosomal aberration and high level mosaic blastocysts with mosaicism levels between 50% and 80%.	Split-sample validation	No	854/284/284 (6:2:2)
V. S. Jiang et al.,[37] 2023	PGT-A (modified FAST-SeqS NGS)	NR	Split-sample validation	No	NR
S. Rajendran et al.,[38] 2023	PGT-A (NGS)	NR	Four-fold cross-validation	Yes	884/295/505
L. Sun et al.,[39] 2024	PGT-A (NGS)	NR	Split-sample validation	Yes	NR
N. Handayani et al.,[40] 2024	PGT-A (NGS)	Euploid (a mixture of euploid and <30% aneuploid cells), aneuploid (mosaicism with more than 80% aneuploid cells), and mosaic (a mixture of euploid and 30–80% aneuploid cells, with low-level mosaicism defined as 30–50% aneuploid cells while the remaining cells were categorized as high-level mosaicism)	Split-sample validation	No	816/NR/204 (8:2)
B. X. Ma et al.,[41] 2024	PGT (NGS)	A threshold of more than 70% was established for the detection of aneuploidy. When it comes to chromosomes, the threshold for mosaic detection differs. The lower limit was 30% for chromosomes 13, 16, 18, and 21, 50% for chromosome 19, and 40% for all other chromosomes. A number that is below the lower bound denotes euploidy.	NR	No	NR
H. He et al.,[42] 2024	PGT-A	NR	10-fold cross-validation	No	NR

Abbreviation: aCGH: array comparative genomic hybridization; NGS: next generation sequencing; NR: not reported; PGT-A: preimplantation genetic testing for aneuploidy; SNP: single nucleotide polymorphism. [a]Studies (n = 12) included in the meta-analysis.

Table 2: **Gold standard, detection thresholds and model validation for the 20 included studies.**

Author, year	Equipment used to acquire imaging data	Perform image pre-processing	Exclusion of a poor-quality image	Images and image-based annotations	Nonimage annotations
T. Bamford et al.,[26] 2023[a]	Time-lapse (EmbryoScope TLS, Vitrolife; Frölunda, Sweden)	Yes	NR	Morphokinetic parameters	Embryologic predictors (genetic platform, IVF or ICSI, sperm concentration, sperm progressive motility), clinical predictors (age of egg provider, number of eggs retrieved, sperm provider age, long or short protocol, FSH dose).
J. Barnes et al.,[27] 2023[a]	Time-lapse (Embryoscope)	Yes	Yes	Static images (500 × 500 pixels) captured at 110 h after ICSI, morphokinetic parameters, blastocyst morphological assessments (blastocyst grade, blastocyst scor).	Maternal age
A. Chavez-Badiola et al.,[28] 2020[a]	Inverted microscopes: Olympus IX71 (laser, 400X, or 200× objectives) (640× 480 pixels) or Olympus IX73 (400X or 200× objectives) (807× 603 pixels) using standard light optics.	Yes	Yes	Static images taken during 5 or 6 days after fertilization before any intervention, such as biopsy, cryopreservation, or transfer.	No
B. Huang et al.,[29] 2021[a]	Time-lapse (Embryoscope Plus, Vitrolife, Denmark)	Yes	Yes	Static images captured during 5 or 6 days after fertilization before biopsy and video files of entire cleavage stage and the blastocyst stage, kinetic parameters, blastocyst stage.	The age of blastocyst (Day5 or Day 6), patient's age.
C. I. Lee et al.,[30] 2021[a]	Time-lapse (EmbryoScope+, Vitrolife, Sweden)	Yes	NR	Time-lapse videos	No
S. De Gheselle et al.,[31] 2022[a]	Time-lapse (EmbryoScope or EmbryoScope Plus; Vitrolife, Sweden)	No	NR	Morphokinetic, standard embryonic development features.	Subjects' demographic and clinical and cycle features (sperm characteristics, woman's age at the start of treatment, and the total dose of gonadotropins).
S. M. Diakiw et al.,[19] 2022[a]	Standard optical light microscopy imaging system	Yes	Yes	Static images were collected of embryos on Day 5, Day 6 and Day 7 of culture, with only Day 5 embryo images used for training and development of the AI model.	No
Y. Zou et al.,[32] 2022[a]	Time-lapse (EmbryoSlide, Vitrolife, Frölunda, Sweden)	Yes	Yes	Morphokinetic parameters, dysmorphisms and irregular cleavages and blastocyst quality.	Clinical features (maternal and paternal age, BMI, basal sex hormone, PGT indications, number of ovarian stimulation cycles, ovarian stimulation protocol, ovarian stimulation days, gonadotrophin dose, number of oocytes and embryos at stimulation day, semen quality and endometrial thickness).
G. B. Danardono et al.,[33] 2023[a]	Time-lapse (MIRI time-lapse incubators) and inverted microscope (Olympus IX71 or Nikon Eclipse Ti, Japan).	Yes	Yes	Static image extraction from time-lapse videos recorded through a closed incubator system and direct image extraction captured using an inverted microscope.	Clinical Characteristics (Etiology of Infertility, basal sex hormone levels, Stimulation Protocol, Estradiol and Progesterone on trigger day, AMH, AFC), baseline (Female Age, BMI, Infertility Duration, Type of Infertility).
E. Paya et al.,[20] 2023[a]	Time-lapse (EmbryoScope or EmbryoScope Plus TL system, Vitrolife, Frölunda, Sweden)	Yes	NR	Videos and static images provided by EmbryoScope time-lapse system with a resolution of 500*500 pixels which were taken automatically every 10–20 min and in up to 7–11 focal planes, morphokinetic parameters.	Female age
Z. Yuan et al.,[18] 2023[a]	Time-lapse (Vitrolife)	No	NR	Morphokinetic parameters, gardner grade.	Female age, frozen days.
T. M. Luong et al.,[34] 2024[a]	Time-lapse (EmbryoScope; Vitrolife, Goteborg, Sweden)	NR	NR	Morphokinetic parameters, morphology grades.	Parental clinical data (maternal age, paternal age, BMI, AMH, LH, E2, P4, oocyte number, IVF number, sperm concentration, and sperm motility on an embryo-by-embryo basis).
J. A. Ortiz et al.,[35] 2023	Optical microscope	No	NR	Embryo quality	Clinical Characteristics (maternal age, paternal age and karyotype, indications for PGT-A, doses of gonadotrophins used, and the number of oocytes retrieved, day of biopsy).
F. Chen et al.,[36] 2023	Time-lapse (Embryoscope or Embryoscope Plus, Vitrolife, Copenhagen, Denmark)	Yes	NR	Time-lapse videos	Female age, male age, AMH, adverse pregnancy events, parental chromosomal structural abnormality, immunological abnormalities, and semen abnormalities.

(Table 3 continues on next page)

Author, year	Equipment used to acquire imaging data	Perform image pre-processing	Exclusion of a poor-quality image	Images and image-based annotations	Nonimage annotations
(Continued from previous page)					
V. S. Jiang et al.,[37] 2023	Time-lapse (EmbryoScope, Vitrolife)	Yes	Yes	Static image collected at 10-min intervals under illumination from a single 635 nm LED using a Leica 20 × objective.	Maternal age, AMH level, paternal sperm quality, total number of normally fertilized (2 PN) embryos.
S. Rajendran et al.,[38] 2023	Time-lapse (Embryoscope or Embryoscope+)	Yes	Yes	Time-lapse sequences typically constituted 360–420 distinct frames, captured at 0.3-h intervals over five days of development, morphological grades and morphokinetic parameters, Blastocyst Score, ICM Score, TE Score, Expansion Score.	Maternal age
L. Sun et al.,[39] 2024	Time-lapse (EVO; Vitrolife Kft, Szeged, Hungary)	Yes	NR	Time-lapse videos or static image extraction from time-lapse videos, morphokinetic parameters.	Clinical metadata (e.g., maternal age, body mass index)
N. Handayani et al.,[40] 2024	Time-lapse (Miri TL; Esco Medical, Denmark)	Yes	NR	Static image extraction from time-lapse videos.	No
B. X. Ma et al.,[41] 2024	Time-lapse (Embryoscope Plus, Vitrolife A/S, Denmark) incubator	No	NR	Time-lapse videos, morphological assessment.	Length of blastocyst incubation, parental chromosome results, embryo's cleavage pattern.
H. He et al.,[42] 2024	Time-lapse (Embryoscope Plus, Vitrolife, Denmark)	No	NR	Morphokinetic parameters	non-invasive chromosomal screening (NICS)

AFC: antral follicle count; AMH: anti-mullerian hormone; BMI: body mass index; E2: estradiol; ICSI: intracytoplasmic sperm injection; IVF: *in vitro* fertilization; LH: luteinizing hormone; NR: not reported; P4: progesterone. [a]Studies (n = 12) included in the meta-analysis.

Table 3: **Equipment, image pre-processing and annotations for the 20 included studies.**

validation approaches, risk of bias, maternal age, geographical location, sample size, and year of publication. Sensitivity and specificity were used as the primary response variables to evaluate model performance. A bivariate normal distribution was assumed for the random error distribution, with a logit link function applied. The random effects term was assumed to follow a normal distribution.[51] Using Scatterplots to confirm the linearity for quantitative predictors. Additionally, publication bias was assessed using Deek's funnel plot asymmetry test,[52] implemented via the MIDAS module in Stata with a p-value of less than 0.05 was considered indicative of publication bias. The clinical applicability of the studies was assessed using a Fagan diagram.

Nine subgroup analyses were conducted to explore sources of heterogeneity: (1) based on the type of AI algorithm (ML vs. DL); (2) stratified by the type of AI-driven DSS (black-box, matte-box, or glass-box); (3) according to annotation methods (image-only vs. image plus clinical data); (4) external validation (with vs. without external validation); (5) by risk of bias (≥3 domains with low risk vs. <3 domains with low risk); (6) inclusion of maternal age (yes vs. no); (7) geographical region (Asia vs. non-Asia); (8) sample size (<400 vs. >400); and (9) publication year (before 2023 vs. after 2023).

In the systematic review, the methodological robustness of each included study was assessed using the QUADAS-AI tool as implemented in Review Manager (RevMan, Version 5.4). To visually illustrate the variance in Se and Sp estimates across studies, a crosshairs plot was generated using R (Version 4.4.0). All additional statistical analyses were performed in STATA (version 17, STATA Corp., College Station, TX, USA) with the MIDAS module[53] and random-effects models and Meta-DiSc 1.4 software,[54] employing a two-tailed significance level set at a type I error probability of 0.05.

Role of funding source

Our study was funded by the National Key R&D Program of China (2022YFC2702905), the Shengjing Freelance Researcher Plan of Shengjing Hospital and the 345 talent project of Shengjing Hospital. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all study data and took final responsibility for the decision to submit the manuscript for publication.

Results

Study selection and characteristics of included studies

In the initial search, a total of 4774 records were identified. Following the removal of 1543 duplicates, the remaining records underwent a title and abstract screening process, which led to the exclusion of 2837 studies. Subsequently, 65 studies were selected for full-text review. Ultimately, 20 articles met the inclusion criteria for this systematic review, and among these, 12 provided sufficient data to be incorporated into the meta-analysis (Fig. 1).

Author, year	Type of AI-driven DSS	Transfer learning applied	Study design, source of data, sample period	Open access data
T. Bamford et al.,[26] 2023[a]	Matte-box or Glass-box	NR	Retrospective multicenter cohort study, data from nine IVF clinics in the UK, 2012–2020.	No
J. Barnes et al.,[27] 2023[a]	Black-box or Matte-box	NR	Retrospective study, data from Weill Cornell Medicine Centre of Reproductive Medicine, New York, NY, USA, 2012–2017.	No
A. Chavez-Badiola et al.,[28] 2020[a]	Black-box	NR	Retrospective study, data from three New Hope Fertility Centres in Mexico City, Guadalajara, and New York City, 2015.1–2019.6.	No
B. Huang et al.,[29] 2021[a]	Matte-box	Yes	Retrospective single-centre cohort study, data from Reproductive Medicine Centre of Tongji Hospital, Huazhong University of Science and Technology, Wuhan, China, 2018.4–2020.12.	No
C. I. Lee et al.,[30] 2021[a]	Black-box	Yes	Retrospective study, data from Division of Infertility, Lee Women's Hospital, Taichung, Taiwan, NR.	No
S. De Gheselle et al.,[31] 2022[a]	Glass-box	No	Retrospective cohort analysis, data from Department for Reproductive Medicine of Ghent University Hospital (Belgium), 2016.01–2019.12.	No
S. M. Diakiw et al.,[19] 2022[a]	Black-box	Yes	Most data were collected retrospectively, with additional data collected prospectively for double-blind evaluation of the final genetics AI model, data from 10 different IVF clinics located in the USA, India, Spain, and Malaysia, 2011–2021.	No
Y. Zou et al.,[32] 2022[a]	Matte-box or Glass-box	Yes	Retrospective study, data from Shanghai Jiai Genetic and Infertility Institute, Obstetrics and Gynecology Hospital of Fudan University, China, 2016.7–2021.7.	No
G. B. Danardono et al.,[33] 2023[a]	Matte-box	Yes	Retrospective study, data from the Morula IVF Jakarta Clinic, Jakarta, Indonesia, NR.	No
E. Paya et al.,[20] 2023[a]	Black-box or Matte-box	Yes	Retrospective study, data from IVI Valencia, Spain, NR.	No
Z. Yuan et al.,[18] 2023[a]	Glass-box	No	Retrospective study, data from Reproductive Medicine Centre of Xuzhou Maternal and Child Health Care Hospital, 2019.01–2022.01.	No
T. M. Luong et al.,[34] 2024[a]	Glass-box	No	Retrospective cohort study, data from Taipei Fertility Centre in Taipei, Taiwan, 2020.03–2022.08.	No
J. A. Ortiz et al.,[35] 2023	Glass-box	No	Retrospective and observational study, data from Instituto Bernabeu, Alicante, Spain, 2013.01–2020.12.	No
F. Chen et al.,[36] 2023	Matte-box	Yes	Retrospective study, data from Reproductive Centre of The First Affiliated Hospital of Sun Yat-sen University, 2020.02–2021.05.	No
V. S. Jiang et al.,[37] 2023	Matte-box	NR	Retrospective study, data from Massachusetts General Hospital Fertility Centre in Boston, Massachusetts, 2019–2022.	No
S. Rajendran et al.,[38] 2023	Matte-box	Yes	NR, data from Weill Cornell Medicine's Centre for Reproductive Medicine, IVI Valencia, Spain, and IVF Florida, USA, 2018–2020.	No
L. Sun et al.,[39] 2024	Black-box or Matte-box	Yes	Most data were collected retrospectively (2010.03–2018.12.31), with additional data collected prospectively (2019–2023) from Guangzhou Women and Children's Hospital and Jiangmen Central Hospital, China.	No
N. Handayani et al.,[40] 2024	Black-box	Yes	Single-centre cohort study, data from private online data-base of Morula IVF Jakarta Clinic, Jakarta, Indonesia, 2021.01–2022.10.	No
B. X. Ma et al.,[41] 2024	Black-box or Matte-box	Yes	Retrospective cohort study, data from Reproductive Medicine Centre, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China, 2018–2021.	No
H. He et al.,[42] 2024	Glass-box	No	Retrospective study, data from Reproductive Medicine Centre at Tongji Hospital, 2020.09–2021.09	No

DSS: Decision Support Systems; NR: not reported. [a]Studies (n = 12) included in the meta-analysis.

Table 4: **Type of AI-driven DSS, study design, source of data and sample period for the 20 included studies.**

The majority of the studies (n = 16) were retrospective, with only two employing prospective data collection for the double-blind evaluation of the final AI model, two remaining studies did not specify the type of research conducted. None of the studies utilized images from open-access databases. Eight studies excluded low-quality images, whereas the remaining twelve did not mention this process. External validation using non-sample datasets was conducted in seven studies. The distribution of research on AI algorithms in this study is as follows: DL was used in ten studies, ML in five, and both DL and ML in five. According to different annotation extraction and ploidy prediction steps, AI-driven DSSs are categorized into three types: 1. Black-box: Refers to AI models that directly make predictions from raw image data without transparency in the decision-making process. The internal workings are not interpretable. 2. Matte-box: It involves an intermediate step where data, either manually or automatically annotated, is input into a black-box model. This approach enhances performance but still lacks interpretability in the final prediction stage. 3. Glass-box: Combines manual or automatic annotation with interpretable ML models in the prediction step. This allows the prediction process to be transparent and explainable, offering insights into how specific decisions are made.[17] The number of studies for each type was black-box (n = 4), matte-box (n = 5), glass-box (n = 5), black-box or matte-box (n = 4), matte-box or glass-box (n = 2). Table 1 through 4 present detailed characteristics of the studies included in the analysis.

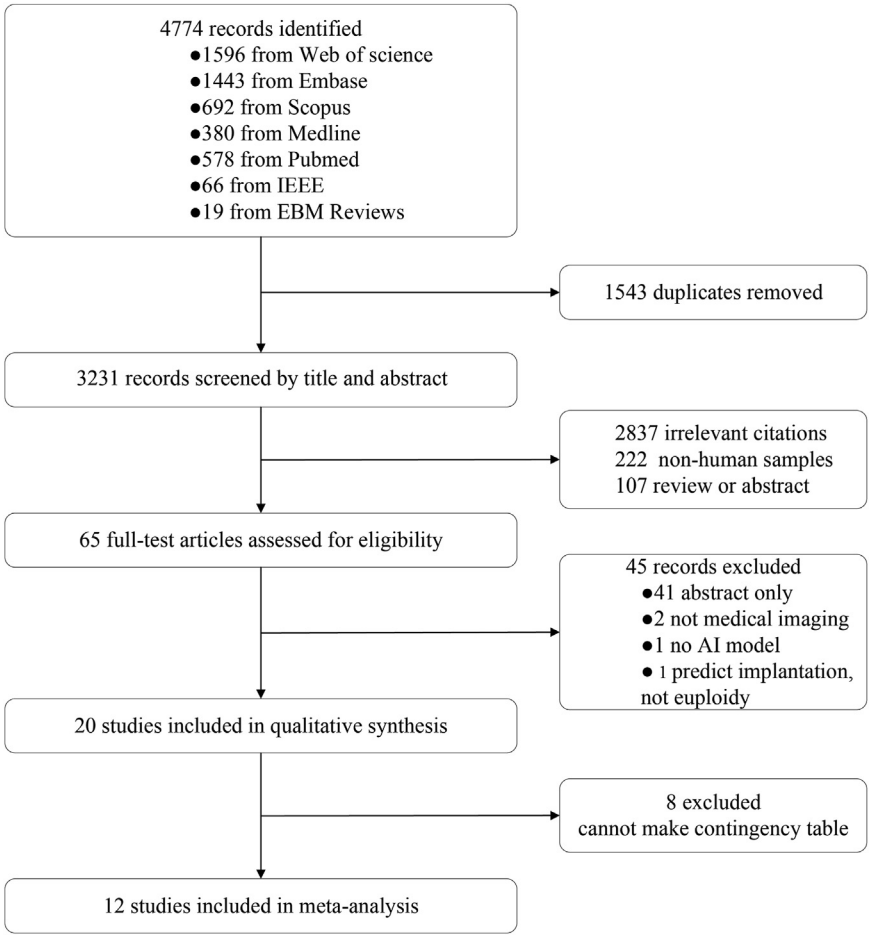

Fig. 1: PRISMA flowchart of study selection.

Pooled performance of AI algorithms

The SROC curves for 12 included studies encompassing 124 contingency tables are provided in Fig. 2a, showing individual studies and summary estimates of diagnostic accuracy, the combined Se and Sp for all AI algorithms were 0.67 (95% CI: 0.64–0.70) and 0.58 (95% CI:0.54–0.61), respectively, with an AUC of 0.67 (95% CI: 0.62–0.71). When selecting the contingency table with the highest accuracy from these studies, the pooled Se and Sp improved to 0.71 (95% CI: 0.59–0.81) and 0.75 (95% CI: 0.69–0.80), respectively, with an AUC of 0.80 (95% CI: 0.76–0.83) (Fig. 2b), indicates that the AI model demonstrates good accuracy, suggesting its potential for clinical application.

Fig. 3 presents a crosshairs plot illustrating the reported point estimates and confidence intervals. This figure integrates elements from both ROC and forest plots to illustrate the bivariate relationship between sensitivity and specificity. It also captures the extent of heterogeneity among studies, as evidenced by the variability in arm lengths and the distribution of data points throughout the plot.[55]

To investigate the clinical utility of AI, a Fagan nomogram was generated. Assuming a 46% prevalence of euploid embryos, the Fagan nomogram shows that the posterior probability of euploid embryos was 71% if the test was positive, and the posterior probability of euploid embryos was 25% if the test was negative (Fig. 4). Consequently, the positive predictive value (PPV) was 71%, and the negative predictive value (NPV) was 75%.

Quality assessment

The quality of the studies included in this analysis was assessed using the QUADAS-AI tool, as shown in Supplementary Fig. S1. Detailed assessment results are depicted in Supplementary Fig. S2. Notably, 19 studies exhibited a high or unclear risk of bias in patient selection, while 13 studies showed a similar risk concerning the index test. This was primarily attributed to

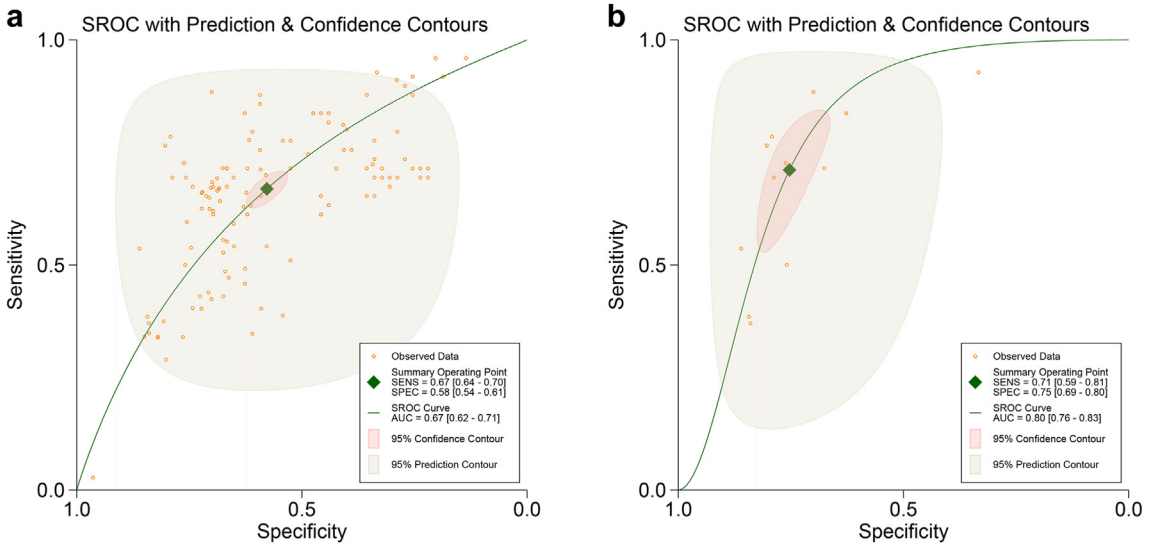

Fig. 2: SROC for sensitivity, specificity and diagnostic accuracy of AI model for prediction of embryonic ploidy. a: SROC curves of all studies included in the meta-analysis (12 studies with 124 tables). b: SROC curves of studies when selecting contingency tables reporting the highest accuracy (12 studies with 12 tables). Abbreviations: SROC: summary receiver operating characteristic; SENS: summary sensitivity; SPEC: summary specificity.

the absence of data from open sources and the lack of rigorous external validation.

Heterogeneity analysis

Heterogeneity was estimated in the forest plot (Supplementary Fig. S21), where sensitivity and specificity exhibited substantial heterogeneity (I^2 = 97.72 (95% CI: 97.08–98.36), p < 0.0001, and I^2 = 92.24 (95% CI: 89.07–95.41), p < 0.0001, respectively). However, no article with a relevant impact on heterogeneity was found using sensitivity analysis (Supplementary Fig. S32).

The threshold effect analysis indicated a significant threshold effect contributing to the observed heterogeneity within this study (Spearman correlation estimate = 0.606, p < 0.001). This suggests that variations in the cutoff values used for diagnosing euploid embryos in PGT-A represent a potential source of heterogeneity in the research findings.

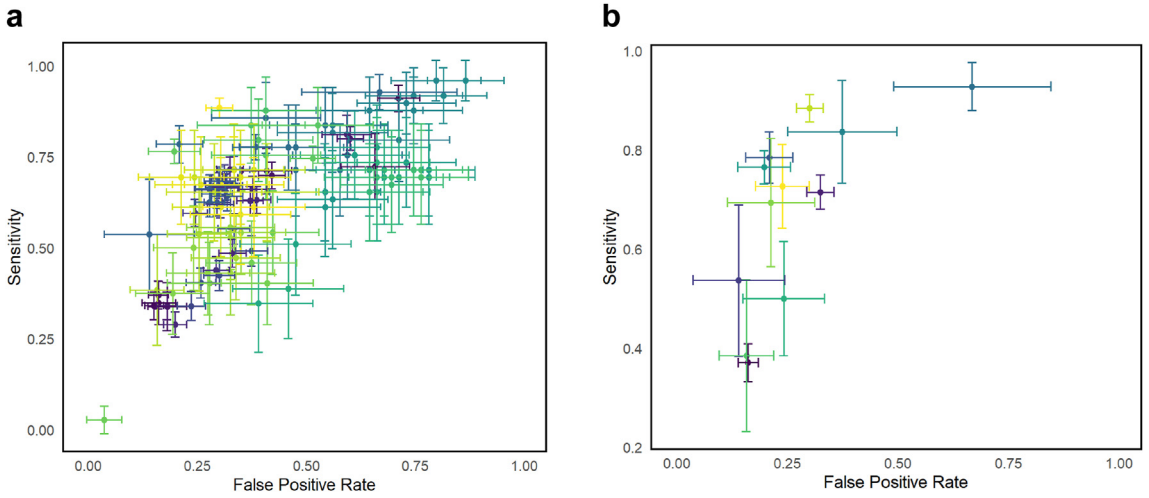

Fig. 3: Cross-hair Plot for sensitivity and false positive rate of AI model for prediction of embryonic ploidy. a: Cross-hair Plot of all studies included in the meta-analysis (12 studies with 124 tables). b: Cross-hair Plot of studies when selecting contingency tables reporting the highest accuracy (12 studies with 12 tables).

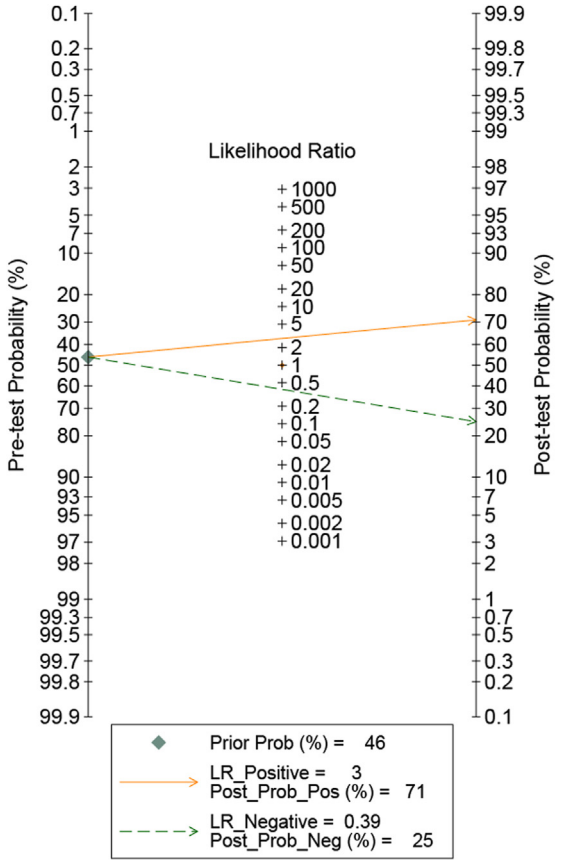

Fig. 4: Fagan normogram for the prediction of euploid embryos based on embryonic images. Abbreviations: Post_Prob_Pos: positive posterior probability; Post_Prob_Neg: negative posterior probability; LR: likelihood ratio.

Multivariable meta-regression was performed to explore the sources of heterogeneity among the studies, with the detailed findings presented in Table 5. The results indicate that Algorithm ($p < 0.001$), type of AI-driven DSS ($p = 0.03$), type of annotation ($p < 0.001$), external validation ($p = 0.02$), risk of bias ($p = 0.01$), maternal age ($p < 0.001$), sample size ($p < 0.001$), and year of publication ($p = 0.01$) contribute to the heterogeneity in sensitivity, while, type of AI-driven DSS ($p < 0.001$), external validation ($p < 0.001$), risk of bias ($p < 0.001$), geographical distribution ($p < 0.001$), and year of publication ($p < 0.001$) are sources of heterogeneity in specificity.

Bivariate boxplot visualizations (Supplementary Fig. S31) were used to illustrate the interdependence and potential negative correlation between sensitivity and specificity. Sensitivity was found to be slightly higher than specificity, consistent with the common inverse relationship observed in diagnostic test accuracy studies.[56] Moreover, Deek's funnel plot asymmetry test indicated no significant evidence of publication bias ($p = 0.85$) (Supplementary Fig. S33).

Subgroup meta-analyses

To further explore the potential sources of heterogeneity, we conducted subgroup meta-analyses stratified by algorithm type, AI-driven DSS categories, annotation methods, model validation techniques, risk of bias, maternal age, geographical region, sample size, and year of publication.

Subgroup analyses revealed that DL models outperformed ML models in terms of AUC (0.71 vs. 0.63). Studies using both image and non-image data demonstrated better predictive performance compared to image-only studies (AUC 0.71 vs. 0.62). External validation and lower risk of bias were associated with more reliable results (AUC 0.70 vs. 0.64 and 0.71 vs. 0.61, respectively), and including maternal age improved model performance (0.71 vs. 0.62). Larger sample sizes generally produced higher specificity and AUC values. Publication year also influenced outcomes, with more recent studies showing improvements in specificity and AUC (Table 5, Supplementary Figs. S3–S11, and S22–S30).

Discussion

Although PGT-A is highly accurate in detecting chromosomal abnormalities and is frequently employed by clinics to enhance pregnancy outcomes, its associated risks remain contentious.[57] Recent evidence indicates that invasive genetic testing may increase the risk of preeclampsia (adjusted OR = 3.02; 95% CI: 1.10–8.29) and placenta previa (adjusted OR = 4.56; 95% CI: 0.93–22.44),[58] while may not significantly improve pregnancy or live birth rates, questioning its clinical utility.[59] Thus, due to the invasive nature of PGT-A and its clinical controversies, there is a need for accurate non-invasive methods to predict embryo ploidy.

AI has been extensively applied across various clinical fields.[60–63] In assisted reproduction, the integration of AI offers a standardized and potentially more objective method for evaluating embryos.[64,65] To our knowledge, this is the first systematic review and meta-analysis focused on using AI to predict embryo ploidy based on imaging data. In alignment with established guidelines for diagnostic reviews,[61] we conducted an exhaustive literature search spanning medical, engineering, and technology databases to ensure methodological rigor and interdisciplinary analysis. In this study, twenty eligible studies were identified with twelve studies included in the meta-analysis. The pooled Se, Sp, and AUC for AI-based prediction of embryonic euploidy were 0.71 (95% CI: 0.59–0.81), 0.75 (95% CI: 0.69–0.80), and 0.80 (95% CI: 0.76–0.83), respectively, based on a total of 6879 embryos (3110 euploid and 3769 aneuploid).

	No. of studies (tables)	Sensitivity			p value[b]	Specificity			p value[b]	AUC
		Sensitivity	p value[a]	I^2 (95%CI)		Specificity	p value[a]	I^2 (95%CI)		
Overall	12 (124)	0.67 (0.63–0.70)	<0.0001	95.71 (95.27–96.15)		0.58 (0.54–0.61)	<0.0001	95.72 (95.28–96.17)		0.67 (0.62–0.71)
Algorithm					<0.001				0.46	
Machine learning	6 (73)	0.69 (0.64–0.73)	<0.0001	94.98 [94.27–95.68)		0.50 (0.44–0.55)	<0.0001	96.34 [95.87–96.81)		0.63 [0.58–0.67]
Deep learning	8 (51)	0.64 (0.60–0.67)	<0.0001	96.37 (95.81–96.92)		0.68 (0.65–0.70)	<0.0001	93.81 (92.70–94.93)		0.71 (0.67–0.75)
Type of AI-driven DSS					0.03				<0.001	
Black-box	5 (21)	0.64 (0.57–0.70)	<0.0001	96.51 (95.67–97.34)		0.68 (0.64–0.71)	<0.0001	88.41 (84.44–92.38)		0.71 (0.67–0.75)
Matte-box	5 (30)	0.64 (0.59–0.68)	<0.0001	96.24 (95.48–97.00)		0.68 (0.64–0.71)	<0.0001	95.36 (94.35–96.36)		0.71 (0.67–0.74)
Glass-box	6 (73)	0.69 (0.64–0.73)	<0.0001	94.98 (94.27–95.68)		0.50 (0.44–0.55)	<0.0001	96.34 (95.87–96.81)		0.63 (0.58–0.67)
Type of annotation					<0.001				0.99	
Image annotation	8 (65)	0.66 (0.60–0.70)	<0.0001	95.08 (94.35–95.81)		0.52 (0.46–0.57)	<0.0001	94.75 (93.95–95.54)		0.62 (0.58–0.66)
Image plus clinical annotation	9 (59)	0.68 (0.64–0.72)	<0.0001	95.69 (95.04–96.33)		0.64 (0.60–0.68)	<0.0001	96.13 (95.57–96.69)		0.71 (0.67–0.75)
External validation					0.02				<0.001	
Yes	5 (43)	0.62 (0.57–0.66)	<0.0001	97.93 (97.64–98.22)		0.68 (0.65–0.72)	<0.0001	97.32 (96.91–97.72)		0.70 (0.66–0.74)
No	7 (81)	0.69 (0.65–0.73)	<0.0001	88.46 (86.48–90.45)		0.51 (0.46–0.56)	<0.0001	91.04 (89.62–92.47)		0.64 (0.60–0.68)
Risk of bias					0.01				<0.001	
Low	6 (56)	0.63 (0.59–0.66)	<0.0001	97.32 (96.97–97.67)		0.68 (0.65–0.71)	<0.0001	96.55 (96.05–97.04)		0.71 (0.66–0.74)
High	6 (68)	0.70 (0.65–0.74)	<0.0001	90.26 (88.53–92.00)		0.48 (0.42–0.53)	<0.0001	91.86 (90.49–93.24)		0.61 (0.57–0.66)
Maternal age					<0.001				0.99	
Yes	9 (59)	0.68 (0.64–0.72)	<0.0001	95.69 (95.04–96.33)		0.64 (0.60–0.68)	<0.0001	96.13 (95.57–96.69)		0.71 (0.67–0.75)
No	8 (65)	0.66 (0.60–0.70)	<0.0001	95.08 (94.35–95.81)		0.52 (0.46–0.57)	<0.0001	94.75 (93.95–95.54)		0.62 (0.58–0.66)
Geographical distribution					0.92				<0.001	
Asia	6 (19)	0.53 (0.40–0.65)	<0.0001	95.98 (94.92–97.03)		0.72 (0.66–0.78)	<0.0001	82.60 (75.56–89.64)		0.70 (0.66–0.74)
Non Asia	6 (105)	0.69 (0.66–0.72)	<0.0001	95.68 (95.20–96.17)		0.55 (0.51–0.59)	<0.0001	96.18 (95.77–96.59)		0.67 (0.62–0.71)
Sample Size					<0.001				0.57	
<400	6 (80)	0.69 (0.65–0.73)	<0.0001	86.29 (83.80–88.78)		0.51 (0.46–0.56)	<0.0001	89.90 (88.22–91.58)		0.64 (0.59–0.68)
≥400	6 (44)	0.63 (0.57–0.67)	<0.0001	98.06 (97.79–98.32)		0.68 (0.65–0.72)	<0.0001	97.25 (96.84–97.67)		0.71 (0.67–0.74)
Year of publication					0.01				<0.001	
Before 2023	6 (68)	0.70 (0.66–0.75)	<0.0001	89.19 (87.20–91.18)		0.47 (0.42–0.52)	<0.0001	90.67 (89.02–92.31)		0.62 (0.58–0.66)
After 2023	6 (56)	0.62 (0.57–0.66)	<0.0001	97.37 (97.02–97.72)		0.68 (0.65–0.71)	<0.0001	96.40 (95.87–96.93)		0.71 (0.66–0.74)

[a]p-value assessing the heterogeneity within each subgroup. [b]p-value assessing the heterogeneity between subgroups with multivariable meta-regression analysis.

Table 5: **Summary estimate of pooled performance of artificial intelligence in Image-based non-invasive prediction of human blastocyst ploidy.**

Although AI algorithms present great potential for predicting embryonic euploidy, the algorithms developed currently lack the accuracy and robustness required to replace PGT-A in clinical settings and still need to be further improved and validated in randomized clinical trials before clinical application, with an ultimate goal of establishing a robust model with high reliability and accuracy to predict embryo ploidy status. At present, a more feasible approach to applying AI in clinical practice is to use it as a decision-support tool, providing a standardized, non-invasive method to optimize the prioritization of biopsied or transferred embryos.[66]

AI algorithms demonstrate promising potential for various applications in the field of reproductive medicine. Nonetheless, these technologies do have their limitations. It is imperative to thoroughly evaluate the following several methodological concerns affecting their efficiency and reliability.

First of all, it is necessary to overcome data limitations and promote standardization in AI training. The efficacy of AI predictive models in clinical applications is fundamentally contingent upon the construction of large and high-quality datasets.[67] In subgroup analysis we detected that studies with a sample size greater than 400 reported an AUC of 0.71 (95% CI: 0.67–0.74), whereas those with a sample size below 400 showed a lower AUC of 0.64 (95% CI: 0.59–0.68), suggesting that larger sample sizes contribute to improved precision and stability of the AUC estimates. Therefore, an adequate sample size is essential for ensuring the accuracy and credibility of diagnostic models. Current challenges include the limited, single-centre training datasets and the lack of standardized image feature annotation, which hinder the broader adoption of AI models. To address this, we propose creating a global network similar to the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative

(IDRI).[68] This network would enable data sharing and identification, facilitating the development of a comprehensive, well-annotated dataset. Additionally, AI models should be trained on large-scale datasets that reflect the demographic, geographic, and disease diversity of patient populations to ensure broad applicability. Ensuring dataset integrity and comprehensiveness is crucial for maximizing AI's potential in the medical field.

Traditional ML algorithms, such as random forests, support vector machines, and regression models, typically necessitate extensive feature engineering, requiring manual extraction and selection of features, and often exhibit poor performance on imbalanced datasets.[69,70] Additionally, labeling complex medical data, such as patient records, can be time-consuming and costly. In contrast, DL models are more flexible, handling unstructured data like images, text, and audio with less reliance on feature engineering. DL models use neural networks that compute weighted sums of inputs across multiple layers, applying nonlinear functions to generate input representations and predict outcomes.[61] However, DL approaches are more prone to overfitting and generally require larger datasets for training.[71] Therefore, combining ML and DL models is recommended to leverage their respective strengths: DL for feature extraction from unstructured data and ML for final predictions on tabular data. This integrated approach enhances data processing, mitigates issues like data imbalance and overfitting, and ensures more robust clinical outcomes.

In addition, integrating mosaicism reporting into AI algorithms for embryo ploidy prediction is of great clinical significance. Many AI models predicting embryo ploidy status are limited by the omission of mosaicism reporting in the algorithms, potentially leading to a loss of vital information and reduced accuracy. The clinical suitability of mosaic embryos for transfer remains debated,[72,73] with studies suggesting that mosaic diagnoses may result from PGT-A amplification methods, biopsy techniques, or poor embryo quality.[73] In fact, many embryos diagnosed as mosaic are later found to be euploid following frozen embryo transfer (FET). Recent studies have demonstrated that mosaic blastocysts exhibit potential for self-correction, leading to successful pregnancies and healthy live births,[74,75] especially in low-level mosaic embryos, which have outcomes similar to euploid embryos.[76] Professional societies recommended prioritizing low-level mosaic transfers when no euploid embryos are available.[77] Given the reproductive potential exhibited by mosaic embryos, future algorithms should contemplate incorporating mosaic embryos in model training and prediction, which could be particularly beneficial in cycles lacking euploid embryos.

External validation, which uses independent datasets to evaluate the reliability and generalizability of diagnostic models across diverse clinical settings, is essential for ensuring their broader adoption.[78,79] Among the 20 studies reviewed, only 7 conducted external validation, indicating a significant gap in understanding model performance in real-world environments. Subgroup analysis showed that studies with external validation had higher diagnostic accuracy (AUC: 0.70, 95% CI: 0.66–0.74) compared to those without (AUC: 0.64, 95% CI: 0.60–0.68). Ramspek et al.[80] also underscore the importance of external validation in evaluating the reproducibility and transportability of predictive models. Therefore, more research incorporating external validation is urgently needed to refine models, enhance diagnostic accuracy, bolster the confidence of healthcare professionals in these models, and ultimately enhance their application and efficacy in clinical decision-making.

In a review of 20 studies, 12 provided sufficient data for establishing contingency tables. Various metrics have been employed to report diagnostic performance in AI research, with Se, Sp, and accuracy being the most commonly used. These metrics are essential for constructing contingency tables that include TP, FP, FN, and TN. Additionally, metrics from computer science, like precision, F1 score, and recall, are sometimes employed. However, these limited data occasionally hinder the construction of comprehensive contingency tables. Many publications fail to effectively communicate their methodologies, often omitting the release of algorithms and datasets, thereby restricting the ability of readers to scrutinize results for errors. To improve replicability and confidence in AI techniques, future research should prioritize sharing raw data and methodologies comprehensively.[69]

To enhance the predictive accuracy of models, several studies incorporated manually annotated morphokinetic parameters and embryo morphology scores.[18,20,26,31,32,35] However, manual annotation is influenced by the researchers' expertise, leading to variability in data interpretation and introducing subjective bias. This may undermine data consistency and limit the generalizability and applicability of AI models..[28] Ideally, AI models should rely on standardized, reproducible data rather than non-standardized subjective metrics. Automatic annotation utilizes AI tools to autonomously label datasets, reducing the time and errors associated with manual annotation, thereby improving data quality, consistency, and model performance. Rajendran et al.[38] applied automatic annotation using Bidirectional Long Short-Term Memory models to assess expansion, inner cell mass, trophectoderm, and overall blastocyst scores. F. Chen et al.[36] developed the AMCFNet model, which autonomously extracted features from clinical data and integrated them with embryonic morphological features. This model demonstrated strong predictive accuracy for identifying euploid blastocysts (AUC = 0.729), assisting embryologists in embryo selection between days 5 and 7. Automatic annotation technology

significantly improves the efficiency and accuracy of embryo image data processing, allowing researchers to effectively select high-quality embryos, which is essential for developing reliable and interpretable AI models.[81,82]

AI-driven DSS for embryonic annotation and ploidy prediction can be categorized into black-box, matte-box, and glass-box models, with increasing levels of interpretability.[17] Interpretability refers to a model's ability to clearly explain its decision-making process in a human-understandable manner. Ensuring model interpretability is critical for fairness and reliability in embryo identification. Currently, two main strategies improve model interpretability. The first integrates clinical parameters, significantly enhancing both interpretability and predictive accuracy.[32] A subgroup analysis of 12 selected studies revealed that AI models based solely on imaging data have limited accuracy in predicting embryonic ploidy (AUC 0.62, 95% CI: 0.58–0.66). However, combining imaging data with clinical annotations improved accuracy (AUC 0.71, 95% CI: 0.67–0.75), highlighting the importance of clinical data integration for predicting euploidy. Moreover, models incorporating maternal age further increased accuracy (AUC 0.71 vs. 0.62), confirming maternal age as a key predictor. La Marca et al. emphasized its role in determining the likelihood and total number of euploid blastocysts.[83] Nonetheless, concerns were raised during the May 2023 ESHRE Journal Club discussion that including maternal age could disproportionately shift model focus toward patient factors rather than embryo-specific characteristics.[84] This emphasizes the need for balancing technical improvements in ML models with clinically relevant variables like female age to optimize embryo ploidy prediction. The second approach utilizes Class Activation Maps (CAM), a key technique in explainable computer vision (XCV). Initially proposed by Zhou et al.,[85] CAM identifies image regions most relevant for category recognition by CNN. It generates heatmaps by projecting CNN output weights onto feature maps from convolutional layers, highlighting areas that significantly influence network decisions. Such technology enhances model interpretability and provides deeper insights into the decision-making processes of DL models.[86] While most studies focus on integrating clinical parameters for interpretability, only one[20] has applied CAM. Future research should expand the use of CAM to optimize model design, improve performance, reduce bias, and strengthen the interpretability and reliability of image and video analysis. In summary, the integration of clinical data and the adoption of innovative approaches are encouraged to improve the interpretability and reliability of models.

Embryo development is a continuous and dynamic process, presenting significant challenges in predicting embryo ploidy. Single time-point images provide limited insight, restricting the predictive power of models. Time-lapse technology enables continuous observation of dynamic embryonic development, but manual review of entire video footage is impractical for embryologists. To address this issue, researchers have introduced optical flow technology, which automatically assesses the dynamic changes in embryonic development by estimating the flow vectors of each pixel in image sequences.[87,88] In this systematic review, only Lee, C.I. et al. utilized optical flow techniques.[30] It is recommended that future research increasingly apply these techniques to the analysis of video data capturing embryo development.

The integration of AI into reproductive medicine poses ethical, patient acceptance, data privacy, and regulatory challenges. Ethical concerns include ensuring informed consent, addressing potential risks to offspring, and clarifying responsibility in the event of errors.[89] Patient acceptance is crucial for successful AI adoption in healthcare, yet current applications often fail to consider patient perspectives. Engaging patients in AI tool design and ensuring transparency may foster trust and broader adoption.[90] AI systems require large amounts of patient data, raising concerns about data privacy, ownership, and protection.[91] Regulatory frameworks, though evolving, remain insufficient to address AI's complexities, particularly concerning its capacity for autonomous learning and real-time adaptation. A move towards global regulatory convergence, beyond the current soft-law approaches, is essential to ensure the safe, ethical, and effective deployment of AI in reproductive medicine.[92] Moreover, in this study, we observed that diagnostic accuracy reported in studies published after 2023 showed an improved AUC of 0.71 (95% CI: 0.66–0.74), compared to an AUC of 0.62 (95% CI: 0.58–0.66) in studies published prior to 2023. This suggests that AI models are rapidly evolving, and we can reasonably expect further improvements in diagnostic accuracy as these models continue to advance over time.

To the best of our knowledge, this is the first systematic review and meta-analysis evaluating the performance of AI in predicting embryo ploidy, with a comprehensive search of relevant studies in databases spanning medicine, engineering, and technology. In addition, no publication bias was detected in the present study, which enhances the reliability while reducing the risk of skewed conclusions by including both positive and negative results. This balanced approach improves the credibility and generalizability of our findings. Of greater importance, we employed QUADAS-AI, a dedicated risk assessment tool designed for AI diagnostic test studies, to critically assess study quality and risk of bias, which is the strength of this systematic review.

It is important to acknowledge the limitations of the present study when interpreting the results. This meta-analysis exhibited significant heterogeneity, which is common in meta-analyses of diagnostic tests due to the inherent difficulty in controlling for all potential

confounding factors. To account for this, we applied a random-effects model, acknowledging the heterogeneity among studies. Furthermore, subgroup analyses and meta-regression were conducted to investigate the sources of the heterogeneity and identified the type of AI-driven DSS, model validation methods, risk of bias, and year of publication as the primary contributors. It is worth noting that the results of this analysis are based on significant heterogeneity, suggesting that these findings may only apply to specific patient populations. There remain practical challenges that need to be addressed before widespread clinical implementation can be considered. Clinicians should take these contextual factors into account when interpreting AI-based predictions of embryonic ploidy. Future research should focus on standardizing methodologies to improve the consistency and broader applicability of AI models in clinical practice.

In addition, this review encompasses studies with limited sample sizes. Insufficient sample sizes may lead to increased risks of overfitting, decreased generalizability, and constraints on model complexity.[93,94] Furthermore, the majority of studies are single-centre and retrospective in nature, which may increase the potential for selection bias. To address these issues, AI developers can employ strategies such as data augmentation, transfer learning, cross-validation, and external validation to enhance model robustness and reliability, thereby mitigating the evaluation errors associated with small sample sizes.[95–97] Future research should prioritize multi-centre, prospective studies to minimize selection bias and improve the generalizability of the models.

In conclusion, this review systematically examined current studies on AI for predicting embryonic ploidy. Our findings indicated that while the current AI models developed cannot entirely replace invasive methods for determining embryo ploidy, AI demonstrates promise as an auxiliary decision-making tool for embryo selection by predicting ploidy, which may help avoid unnecessary biopsies. Furthermore, we advocate for the development and integration of extensive databases, and the conduct of large-sample, multicentre, prospective studies to facilitate the clinical application of AI. Healthcare professionals should become familiar with AI concepts, metrics, and potential applications, embracing the increasing integration of AI into modern medicine.

Contributors

XX, S-SW, Y-JM, NB, and J-CT conceptualized and designed the study. XX, S-SW, and Y-JM conducted the literature search and data extraction. Risk of bias evaluation was performed by XX, MG, and XH. XX, H-LX, S-SW, SG and Y-JM contributed to data analysis and interpretation. XX, S-SW, H-LX, Y-JM, NB, MG, XH, S-WZ, X-YZ, J-RQ, X-DZ and J-CT drafted and edited the manuscript. All authors read and approved the final version of the manuscript, and ensure it is the case. XX, S-SW, and H-LX have independently verified the underlying data to ensure its accuracy. XX, S-SW, H-LX, and Y-JM contributed equally to this work.

Data sharing statement

The search strategy was shown in Supplementary Note S1, and the contingency tables of 12 studies included in the meta-analysis were shown in Supplementary Table S1. The results of risk of bias and publication bias were separately provided in Supplementary Figs. S2 and S33. Additional data are available on request.

Declaration of interests

All authors declare no competing interests.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFC2702905), the Shengjing Freelance Researcher Plan of Shengjing Hospital and the 345 talent project of Shengjing Hospital.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102897.

References

1 Scott RT Jr, Ferry K, Su J, Tao X, Scott K, Treff NR. Comprehensive chromosome screening is highly predictive of the reproductive potential of human embryos: a prospective, blinded, nonselection study. *Fertil Steril*. 2012;97(4):870–875.
2 Elizabeth S, Leonardo P, Almena L-L, Dinorah H-M, Esther L-B. Aneuploidy rates inversely correlate with implantation during in vitro fertilization procedures: in favor of PGT. In: Israel G, ed. *Modern medical genetics and genomics*. Rijeka: IntechOpen; 2018. Ch. 3.
3 Ata B, Kaplan B, Danzer H, et al. Array CGH analysis shows that aneuploidy is not related to the number of embryos generated. *Reprod Biomed Online*. 2012;24(6):614–620.
4 Franasiak JM, Forman EJ, Hong KH, et al. The nature of aneuploidy with increasing age of the female partner: a review of 15,169 consecutive trophectoderm biopsies evaluated with comprehensive chromosomal screening. *Fertil Steril*. 2014;101(3):656–663.e1.
5 Munné S, Blazek J, Large M, et al. Detailed investigation into the cytogenetic constitution and pregnancy outcome of replacing mosaic blastocysts detected with the use of high-resolution next-generation sequencing. *Fertil Steril*. 2017;108(1):62–71.e8.
6 Munné S, Kaplan B, Frattarelli JL, et al. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. *Fertil Steril*. 2019;112(6):1071–1079.e7.
7 Homer HA. Preimplantation genetic testing for aneuploidy (PGT-A): the biology, the technology and the clinical outcomes. *Aust N Z J Obstet Gynaecol*. 2019;59(2):317–324.
8 Ginoza MEC, Isasi R. Regulating preimplantation genetic testing across the world: a comparison of international policy and ethical perspectives. *Cold Spring Harbor perspectives in medicine*. 2020;10(5):a036681.
9 Theobald R, SenGupta S, Harper J. The status of preimplantation genetic testing in the UK and USA. *Hum Reprod*. 2020;35(4):986–998.
10 Ishchuk MA, Lesik EA, Sagurova YM, Komarova EM. TIME-LAPSE technology in modern embryological practice. *JOWD*. 2023;72(6):193–201 [%J JOWD].
11 Liu Y, Qi F, Matson P, et al. Between-laboratory reproducibility of time-lapse embryo selection using qualitative and quantitative parameters: a systematic review and meta-analysis. *J Assist Reprod Genet*. 2020;37(6):1295–1302.
12 Bamford T, Barrie A, Montgomery S, et al. Morphological and morphokinetic associations with aneuploidy: a systematic review and meta-analysis. *Hum Reprod Update*. 2022;28(5):656–686.
13 Bamford T, Smith R, Young S, et al. A comparison of morphokinetic models and morphological selection for prioritizing euploid embryos: a multicentre cohort study. *Hum Reprod*. 2024;39(1):53–61.
14 Riegler MA, Stensen MH, Witczak O, et al. Artificial intelligence in the fertility clinic: status, pitfalls and possibilities. *Hum Reprod*. 2021;36(9):2429–2442.
15 Luong TM, Le NQK. Artificial intelligence in time-lapse system: advances, applications, and future perspectives in reproductive medicine. *J Assist Reprod Genet*. 2024;41(2):239–252.

16 Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563–577.
17 Lee T, Natalwala J, Chapple V, Liu Y. A brief history of artificial intelligence embryo selection: from black-box to glass-box. *Hum Reprod*. 2024;39(2):285–292.
18 Yuan Z, Yuan M, Song X, Huang X, Yan W. Development of an artificial intelligence based model for predicting the euploidy of blastocysts in PGT-A treatments. *Sci Rep*. 2023;13(1):2322.
19 Diakiw SM, Hall JMM, VerMilyea MD, et al. Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. *Hum Reprod*. 2022;37(8):1746–1759.
20 Paya E, Pulgarín C, Bori L, Colomer A, Naranjo V, Meseguer M. Deep learning system for classification of ploidy status using time-lapse videos. *F&S Sci*. 2023;4(3):211–218.
21 Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol*. 1996;49(11):1225–1231.
22 Afnan MAM, Liu Y, Conitzer V, et al. Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Hum Reproduct Open*. 2021;2021(4):hoab040.
23 Wang S, Zhang Y, Lei S, et al. Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: a systematic review and meta-analysis of diagnostic test accuracy. *Eur J Endocrinol*. 2020;183(1):41–49.
24 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev*. 2021;10(1):89.
25 Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
26 Bamford T, Easter C, Montgomery S, et al. A comparison of 12 machine learning models developed to predict ploidy, using a morphokinetic meta-dataset of 8147 embryos. *Hum Reprod*. 2023;38(4):569–581.
27 Barnes J, Brendel M, Gao VR, et al. A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study. *Lancet Digit Health*. 2023;5(1):e28–e40.
28 Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reprod Biomed Online*. 2020;41(4):585–593.
29 Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can predict embryo ploidy status based on time-lapse data. *Reprod Biol Endocrinol*. 2021;19(1):185.
30 Lee CI, Su YR, Chen CH, et al. End-to-end deep learning for recognition of ploidy status using time-lapse videos. *J Assist Reprod Genet*. 2021;38(7):1655–1663.
31 De Gheselle S, Jacques C, Chambost J, et al. Machine learning for prediction of euploidy in human embryos: in search of the best-performing model and predictive features. *Fertil Steril*. 2022;117(4):738–746.
32 Zou Y, Pan Y, Ge N, et al. Can the combination of time-lapse parameters and clinical features predict embryonic ploidy status or implantation? *Reprod Biomed Online*. 2022;45(4):643–651.
33 Danardono GB, Handayani N, Louis CM, et al. Embryo ploidy status classification through computer-assisted morphology assessment. *AJOG Glob Rep*. 2023;3(3):100209.
34 Luong TM, Ho NT, Hwu YM, et al. Beyond black-box models: explainable AI for embryo ploidy prediction and patient-centric consultation. *J Assist Reprod Genet*. 2024;41(9):2349–2358.
35 Ortiz JA, Morales R, Lledó B, et al. Application of machine learning to predict aneuploidy and mosaicism in embryos from in vitro fertilization cycles. *AJOG Glob Rep*. 2022;2(4):100103.
36 Chen F, Xie X, Cai D, et al. Knowledge-embedded spatio-temporal analysis for euploidy embryos identification in couples with chromosomal rearrangements. *Chin Med J*. 2024;137(6):694–703.
37 Jiang VS, Kandula H, Thirumalaraju P, et al. The use of voting ensembles to improve the accuracy of deep neural networks as a non-invasive method to predict embryo ploidy status. *J Assist Reprod Genet*. 2023;40(2):301–308.
38 Rajendran S, Brendel M, Barnes J, et al. Automatic ploidy prediction and quality assessment of human blastocyst using time-lapse imaging. *bioRxiv*. 2023. https://doi.org/10.1101/2023.08.31.555741.
39 Sun L, Li J, Zeng S, et al. Artificial intelligence system for outcome evaluations of human in vitro fertilization-derived embryos. *Chin Med J (Engl)*. 2024;137(16):1939–1949.
40 Handayani N, Danardono GB, Boediono A, et al. Improving deep learning-based algorithm for ploidy status prediction through combined U-NET blastocyst segmentation and sequential time-lapse blastocysts images. *J Reprod Infertil*. 2024;25(2):110–119.
41 Ma BX, Zhao GN, Yi ZF, Yang YL, Jin L, Huang B. Enhancing clinical utility: deep learning-based embryo scoring model for non-invasive aneuploidy prediction. *Reprod Biol Endocrinol*. 2024;22(1):58.
42 He H, Wu L, Chen Y, et al. A novel non-invasive embryo evaluation method (NICS-Timelapse) with enhanced predictive precision and clinical impact. *Heliyon*. 2024;10(9):e30189.
43 Xu HL, Gong TT, Liu FH, et al. Artificial intelligence performance in image-based ovarian cancer identification: a systematic review and meta-analysis. *EClinicalMedicine*. 2022;53:101662.
44 Xue P, Wang J, Qin D, et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. *NPJ Digit Med*. 2022;5(1):19.
45 Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med*. 2021;27(10):1663–1665.
46 Diaz M. Performance measures of the bivariate random effects model for meta-analyses of diagnostic accuracy. *Comput Stat Data Anal*. 2015;83:82–90.
47 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557–560.
48 Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med*. 1988;7(8):889–894.
49 Dwamena B. *Meta-analytical integration of diagnostic accuracy studies in Stata*. 2007.
50 Thabane L, Mbuagbaw L, Zhang S, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Med Res Methodol*. 2013;13(1):92.
51 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20(19):2865–2884.
52 Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58(9):882–893.
53 Dwamena B. *MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies*. Statistical Software Components; 2007.
54 Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol*. 2006;6(1):31.
55 Phillips B, Stewart LA, Sutton AJ. 'Cross hairs' plots for diagnostic meta-analysis. *Res Synth Methods*. 2010;1(3–4):308–315.
56 Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina (Kaunas, Lithuania)*. 2021;57(5):503.
57 Patounakis G, Hill MJ. The preimplantation genetic testing debate continues: first the hype, then the tension, now the hypertension? *Fertil Steril*. 2019;112(2):233–234.
58 Zhang WY, von Versen-Höynck F, Kapphahn KI, Fleischmann RR, Zhao Q, Baker VL. Maternal and neonatal outcomes associated with trophectoderm biopsy. *Fertil Steril*. 2019;112(2):283–290.e2.
59 Cornelisse S, Zagers M, Kostova E, Fleischer K, van Wely M, Mastenbroek S. Preimplantation genetic testing for aneuploidies (abnormal number of chromosomes) in in vitro fertilisation. *Cochrane Database Syst Rev*. 2020;9(9):Cd005291.
60 Brownstein JS, Rader B, Astley CM, Tian H. Advances in artificial intelligence for infectious-disease surveillance. *N Engl J Med*. 2023;388(17):1597–1607.
61 Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*. 2023;186(8):1772–1791.
62 Lång K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol*. 2023;24(8):936–944.
63 Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. *Artif Intell Med*. 2009;46(1):5–17.

64 Cimadomo D, Chiappetta V, Innocenti F, et al. Towards automation in IVF: pre-clinical validation of a deep learning-based embryo grading system during PGT-A cycles. *J Clin Med*. 2023;12(5):1806.
65 Salih M, Austin C, Warty RR, et al. Embryo selection through artificial intelligence versus embryologists: a systematic review. *Hum Reproduct Open*. 2023;2023(3):hoad031.
66 Loewke K, Cho JH, Brumar CD, et al. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. *Fertil Steril*. 2022;117(3):528–535.
67 Kulkarni S, Seneviratne N, Baig MS, Khan AHA. Artificial intelligence in medicine: where are we now? *Acad Radiol*. 2020;27(1):62–70.
68 Armato SG 3rd, McLennan G, Bidaut L, et al. The Lung image database Consortium (LIDC) and image database Resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys*. 2011;38(2):915–931.
69 Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med*. 2018;284(6):603–619.
70 Litjens G, Ciompi F, Wolterink JM, et al. State-of-the-Art deep learning in cardiovascular image analysis. *JACC Cardiovasc Imaging*. 2019;12(8 Pt 1):1549–1565.
71 Manco L, Maffei N, Strolin S, Vichi S, Bottazzi L, Strigari L. Basic of machine learning and deep learning in imaging for medical physicists. *Phys Med*. 2021;83:194–205.
72 Scott RT Jr, Galliano D. The challenge of embryonic mosaicism in preimplantation genetic screening. *Fertil Steril*. 2016;105(5):1150–1152.
73 Popovic M, Dhaenens L, Boel A, Menten B, Heindryckx B. Chromosomal mosaicism in human blastocysts: the ultimate diagnostic dilemma. *Hum Reprod Update*. 2020;26(3):313–334.
74 Lee CI, Cheng EH, Lee MS, et al. Healthy live births from transfer of low-mosaicism embryos after preimplantation genetic testing for aneuploidy. *J Assist Reprod Genet*. 2020;37(9):2305–2313.
75 Lin PY, Lee CI, Cheng EH, et al. Clinical outcomes of single mosaic embryo transfer: high-level or low-level mosaic embryo, does it matter? *J Clin Med*. 2020;9(6):1695.
76 Abhari S, Kawwass JF. Pregnancy and neonatal outcomes after transfer of mosaic embryos: a review. *J Clin Med*. 2021;10(7):1369.
77 Leigh D, Cram DS, Rechitsky S, et al. PGDIS position statement on the transfer of mosaic embryos 2021. *Reprod Biomed Online*. 2022;45(1):19–25.
78 Potash E, Ghani R, Walsh J, et al. Validation of a machine learning model to predict childhood lead poisoning. *JAMA Netw Open*. 2020;3(9):e2012734.
79 Groot OQ, Bindels BJJ, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop*. 2021;92(4):385–393.
80 Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49–58.
81 From time to space: automatic annotation of unmarked traffic scene based on trajectory data. In: Ma H, Wang Y, Xiong R, eds. *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 2018.
82 Feyeux M, Reignier A, Mocaer M, et al. Development of automated annotation software for human embryo morphokinetics. *Hum Reprod*. 2020;35(3):557–564.
83 La Marca A, Capuzzo M, Longo M, et al. The number and rate of euploid blastocysts in women undergoing IVF/ICSI cycles are strongly dependent on ovarian reserve and female age. *Hum Reprod*. 2022;37(10):2392–2401.
84 Serdarogullari M, Liperis G, Sharma K, et al. Unpacking the artificial intelligence toolbox for embryo ploidy prediction. *Hum Reprod*. 2023;38(12):2538–2542.
85 Learning deep features for discriminative localization. In: Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, eds. *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. 2016.
86 APTJA Minh. *Overview of Class activation maps for visualization explainability*. 2023. abs/2309.14304.
87 Hashmi A, Tlili S, Perrin P, et al. Cell-state transitions and collective cell movement generate an endoderm-like region in gastruloids. *Elife*. 2022;11:e59371.
88 Shah STH, Xuezhi X. Traditional and modern strategies for optical flow: an investigation. *SN Appl Sci*. 2021;3(3):289.
89 Rolfes V, Bittner U, Gerhards H, et al. Artificial intelligence in reproductive medicine - an ethical perspective. *Geburtshilfe Frauenheilkd*. 2023;83(1):106–115.
90 Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. *J Patient Center Res Rev*. 2024;11(1):51–62.
91 Tretter M, Ott T, Dabrock P. *AI-produced certainties in health care: current and future challenges*. AI and Ethics; 2023.
92 Palaniappan K, Lin EYT, Vogel S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. *Healthcare (Basel, Switzerland)*. 2024;12(5):562.
93 Rajput D, Wang W-J, Chen C-C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinf*. 2023;24(1):48.
94 Benkendorf DJ, Hawkins CP. Effects of sample size and network depth on a deep learning approach to species distribution modeling. *Ecol Inf*. 2020;60:101137.
95 Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data*. 2021;8(1):53.
96 Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. *Array*. 2022;16:100258.
97 Safonova A, Ghazaryan G, Stiller S, Main-Knorn M, Nendel C, Ryo M. Ten deep learning techniques to address small data problems with remote sensing. *Int J Appl Earth Obs Geoinf*. 2023;125:103569.