Wang *et al.*
*Reproductive Biology and Endocrinology* (2024) 22:59
https://doi.org/10.1186/s12958-024-01232-8

Reproductive Biology and Endocrinology

METHODOLOGY

Open Access

# Testing the generalizability and effectiveness of deep learning models among clinics: sperm detection as a pilot study

Jiaqi Wang[1†], Yufei Jin[1†], Aojun Jiang[2], Wenyuan Chen[2], Guanqiao Shan[2], Yifan Gu[3,4], Yue Ming[5], Jichang Li[5], Chunfeng Yue[6], Zongjie Huang[6], Clifford Librach[7], Ge Lin[3,4], Xibu Wang[8], Huan Zhao[8*], Yu Sun[2,9,10,11*] and Zhuoran Zhang[1*]

## Abstract

**Background** Deep learning has been increasingly investigated for assisting clinical in vitro fertilization (IVF). The first technical step in many tasks is to visually detect and locate sperm, oocytes, and embryos in images. For clinical deployment of such deep learning models, different clinics use different image acquisition hardware and different sample preprocessing protocols, raising the concern over whether the reported accuracy of a deep learning model by one clinic could be reproduced in another clinic. Here we aim to investigate the effect of each imaging factor on the generalizability of object detection models, using sperm analysis as a pilot example.

**Methods** Ablation studies were performed using state-of-the-art models for detecting human sperm to quantitatively assess how model precision (false-positive detection) and recall (missed detection) were affected by imaging magnification, imaging mode, and sample preprocessing protocols. The results led to the hypothesis that the richness of image acquisition conditions in a training dataset deterministically affects model generalizability. The hypothesis was tested by first enriching the training dataset with a wide range of imaging conditions, then validated through internal blind tests on new samples and external multi-center clinical validations.

**Results** Ablation experiments revealed that removing subsets of data from the training dataset significantly reduced model precision. Removing raw sample images from the training dataset caused the largest drop in model precision, whereas removing 20x images caused the largest drop in model recall. by incorporating different imaging and sample preprocessing conditions into a rich training dataset, the model achieved an intraclass correlation coefficient (ICC) of 0.97 (95% CI: 0.94-0.99) for precision, and an ICC of 0.97 (95% CI: 0.93-0.99) for recall. Multi-center clinical validation showed no significant differences in model precision or recall across different clinics and applications.

**Conclusions** The results validated the hypothesis that the richness of data in the training dataset is a key factor impacting model generalizability. These findings highlight the importance of diversity in a training dataset for model

[†]Jiaqi Wang and Yufei Jin contributed equally to this work.

*Correspondence:
Huan Zhao
dahuan4302@163.com
Yu Sun
yu.sun@utoronto.ca
Zhuoran Zhang
zhangzhuoran@cuhk.edu.cn
Full list of author information is available at the end of the article

evaluation and suggest that future deep learning models in andrology and reproductive medicine should incorporate comprehensive feature sets for enhanced generalizability across clinics.

**Keywords** Semen analysis, Sperm detection, Generalizability, Multicenter validation, Deep learning

## Introduction

Deep learning has been increasingly applied to facilitate diagnosis and treatment of various diseases [1, 2]. Taking infertility as an example, which affects one in six couples worldwide [3, 4], numerous deep learning models have been developed with the aim of improving clinical outcomes and optimizing the operational efficiency in in vitro fertilization (IVF) clinics [5–8]. Most of these models take images as input, for instance, to evaluate sperm motility, concentration, and morphology for selecting high-quality sperm for fertilization [9–11] or for diagnosing male infertility [12–14], to help identify and distinguish sperm and debris in testicular sperm samples [15, 16], or to examine the quality of oocytes [17]. Models have also been developed to use embryo images or time-lapse videos to grade embryos [18, 19] and to predict treatment outcomes such as implantation [20], pregnancy [21], and live birth [22–24].

Despite the potential of deep learning models for advancing clinical practice, existing studies focused on improving model accuracy [25–28] or precision [29–32] while little attempt has been made to investigate model generalizability, an essential aspect for deploying deep learning models for clinical applications. Translating a technique from technical development to clinical deployment can involve various factors that impact the generalizability of the developed technique. Regardless of applications or the types of cells to analyze, the first technical step for deep learning models is often to visually identify and locate an object (oocyte [33, 34], sperm [35–39], and embryo [20, 40–42]) in images. Different clinics, however, use different image acquisition conditions (e.g., microscope brands and models, imaging modes [43–45], magnifications [9, 33], illumination intensity, and camera resolutions [13–15, 39] etc.), as evident in Table 1. In addition, even though the images are acquired under the same conditions, sample preprocessing protocols may also be different among clinics (e.g., for sperm analysis using raw semen versus washed samples). These factors inevitably change the appearance of the images for analysis by deep learning models, thus raising concerns over whether the accuracy of a model reported in one clinic could be reproduced in another clinic.

This question is important but has not been investigated in literature. Existing studies [12–15, 35–39, 43–45], were retrospective studies where a retrospectively collected dataset was split into training, validation, and testing sub-datasets. Although such datasets may include data from multiple clinics [10, 11], model validation and testing were still performed under the same data collection conditions as the training dataset. The lack of prospective model validation and testing with new data beyond the retrospectively collected dataset challenges the generalizability of the developed model under different clinical setups. To address this question, what is needed is prospective validation and testing of model generalizability. However, existing studies mainly use accuracy or precision as the sole metric for evaluating the developed models. Reproducibility metrics such as coefficient of variation or intraclass correlation coefficient (ICC) has rarely been reported in literature.

Technically, applying a pre-trained model in different clinics may involve domain shift, that is, the data (in each clinic) used to evaluate the model is drawn from a population different from the training data. Despite the importance and implications of domain shift in image analysis and deep learning have been discussed in the literature [49–53], few studies have explored the specific factors that contribute to domain shift and its impact on model generalizability in real-world clinical settings. For instance, variations in clinical imaging settings – such as differences in microscope models, imaging modes, magnifications, and sample preparation protocol – may all lead to different data distributions between the testing dataset and training dataset. However, it remains unclear how each specific factor/variation affects model generalizability in clinical settings.

Here we fill this knowledge gap by performing ablation studies which quantitatively revealed how model precision and recall were affected by imaging magnification, imaging mode, and sample preprocessing protocols. The workflow of this manuscript is shown in Fig. 1. As a pilot study, we evaluated performance of state-of-the-art deep learning models for detecting and identifying human sperm, due to their wide applications in andrology laboratories and IVF clinics. Based on the ablation studies, we hypothesized that improving the diversity and richness of the training dataset could increase model generalizability. This hypothesis was first tested by calculating the model's ICC for repeated measurements on new samples. Then the hypothesis was prospectively tested via external validation in three clinics (excluding the academic lab where the model was trained) that used different image acquisition conditions and sample preprocessing protocols.

**Table 1** Summary of clinical applications of object detection models in IVF

| Object | Clinical Application | Algorithm | Datasets | | | | Reference |
|---|---|---|---|---|---|---|---|
| | | | Sources | Imaging mode | Resolution | Magnification | |
| **Sperm** | Selecting high-quality sperm during intracytoplasmic sperm injection (ICSI) treatment | YOLO | Single center | Bright field | 128×128 | 60×, 40× | [9] |
| | | VGG | Multi-center | Bright field | 131×131 | 10× | [10] |
| | | VGG | Multi-center | Bright field | 131×131 | 10× | [11] |
| | | YOLO | Single center | Bright field | / | 60× | [46] |
| | Detecting sperm in semen quality analysis for male infertility diagnosis (locating sperm for subsequent measurement of sperm concentration, motility, and morphology) | YOLO | Single center | Phase contrast[a] | 640×480 | 40× | [12] |
| | | YOLO | Single center | Phase contrast | 1280×960 | 10× | [13] |
| | | YOLO | Single center | Phase contrast | 640×480 | 40× | [14] |
| | | YOLO | Single center | Phase contrast | 640×480 | 40× | [43] |
| | | YOLO | Single center | Phase contrast | 640×480 | 40× | [44] |
| | | YOLO | Single center | Hoffman modulation contrast[b] | 448×448 | 40× | [45] |
| | | YOLO | Single center | Hoffman modulation contras | 1664×1664 | / | [35] |
| | | YOLO | Single center | / | / | / | [36] |
| | | YOLO | Single center | Bright field | 640×640 | 10× | [37] |
| | | YOLO VGG | Single center | Bright field | 598×528 | 20× | [38] |
| | | VGG | Single center | Bright field | 150×150 | 40× | [39] |
| | | CNN | Single center | DIC[c] | / | 20×,100× | [47] |
| | Searching for sperm in testicular sperm extraction samples for azoospermia patients | YOLO | Single center | DIC | 3264×2448 1920×1940 | 63× | [15] |
| | | U-Net | Single center | Bright-field Fluorescence | 256×256 | 10× | [16] |
| **Oocyte** | Detecting oocytes for the selection of high-quality oocytes during ICSI | DeepLabV3 | Single center | Bright field | 1392×1024 | 20× | [17] |
| | | U-Net | Single center | Bright field | 1280×1024 | 4×,15×30×,40× | [33] |
| | | CNN | Single center | Bright field | 250×250 | 20× | [34] |
| **Embryo** | Locating embryos for grading and selecting high-quality embryos for transfer | ResNet | Single center | Bright field | 720×480 | / | [18] |
| | | CNN | Single center | Bright field | 250×250 | 20× | [20] |
| | | YOLO | Single center | Bright field | 500×500 | / | [40] |
| | | VGG | Single center | Bright field | / | / | [41] |
| | | AlexNet | Single center | Bright field | / | / | [42] |
| | | EfficientNetV2 | Single center | Bright field | 1024×768 | / | [48] |

[a] Phase contrast is a technique that enhances the contrast of transparent and colorless specimens by converting phase shifts in light passing through the specimen into changes in intensity

[b] Hoffman modulation contrast (HMC) enhances the contrast of unstained transparent samples by modulating the phase and amplitude of transmitted light. It is commonly used for visualizing sperm and oocytes during in vitro fertilization treatment

[c] Differential Interference Contrast (DIC) microscopy is an optical imaging technique that uses polarized light to produce high-contrast images of transparent specimens, enhancing the three-dimensional appearance of structures by exaggerating differences in optical paths

The results validated the hypothesis that the richness of data in the training dataset is a key factor impacting modelgeneralizability.

## Results

### Investigating factors that impact model generalizability

Deep learning is a data-driven approach, and the training dataset deterministically affects model performance. Considering that different clinics use different imaging conditions, we first investigated how model generalizability is affected by imaging magnification, sample preprocessing protocols, and imaging mode. Ablation study was performed where the training images for each factor was removed from the training dataset, then the model was re-trained to compare performance (Supplementary Table 1 and Supplementary Table 2). Model performance was evaluated by model precision and recall. A lower precision indicates a higher rate of false positive detection, and a lower recall indicates a higher rate of missed detection.

**Imaging magnification:** when 20× sperm images were removed from the training dataset (i.e., training the

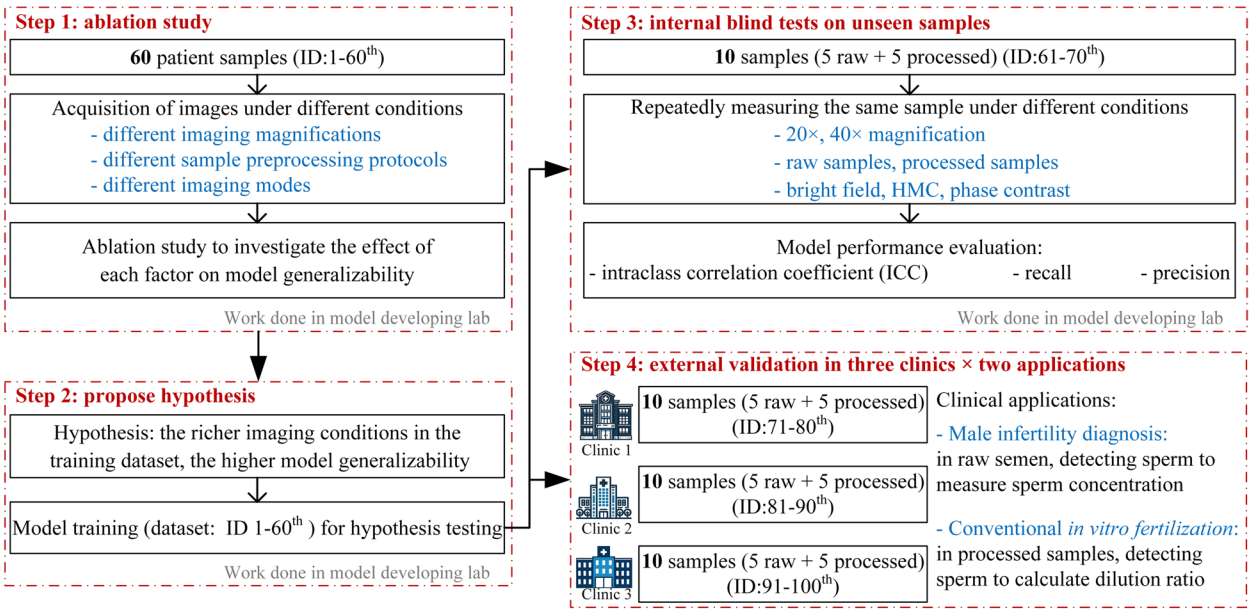


**Fig. 1** Flowchart of study design, including experiments and samples used in each section

model with only 40× sperm images, but testing it with both 20× and 40× images), model precision significantly dropped from 90.64% to 75.09% ($p < 0.01$, Fig. 2A). Model recall also significantly dropped from 92.08% to 15.27% ($p < 0.0001$, Fig. 2A). A higher drop was observed in model recall than precision, possibly because the model learned sperm features from 40× images, and the model perceptual field cannot be mapped directly to 20× images. This interpretation was confirmed by the model weight heatmaps in Fig. 2A. The model raised less weight/attention to sperm, leading to missed detection (drop in recall).

**Sample preprocessing protocols:** when images of raw semen samples were removed from the training dataset, model precision significantly dropped by 58.11% ($p < 0.0001$, Fig. 2B). Raw semen samples contained a high number of non-sperm impurities (e.g., epithelial cells, spermatocytes and leucocytes). Using only processed samples in the training dataset, the ratio between foreground (sperm) and background objects (non-sperm impurities) decreased, making the model to learn features mainly from the sperm but not enough features to distinguish the impurities. As a result, the model falsely raised more weight/attention to impurities and detected them as sperm, leading to a low precision. No significant drop in model recall was observed. This is reasonable because impurities in raw semen does not change the appearance of sperm itself, thus not causing missed detection.

**Imaging mode:** interestingly, we also noticed that when removing Hoffman modulation contrast images from the training dataset, model precision and recall also dropped (Fig. 2C). Although the drops in precision ($p < 0.01$) and recall ($p < 0.1$) are still significant, they are smaller than that caused by removing 20× images or raw sample images. The situation was similar for removing phase contrast images, where model precision and recall dropped by 15.01% ($p < 0.01$) and 15.06% ($p < 0.01$) respectively (Fig. 2D). Hoffman modulation contrast and phase contrast imaging modes mainly changed image contrast, and the resulting images were largely similar to brightfield images. Among the two experiments, the

(See figure on next page.)

**Fig. 2** Ablation studies were performed to investigate how model generalizability is affected by imaging magnification, imaging mode, and sample preprocessing protocols. **A**-**D** In the ablation experiment, each investigated factor was removed from the training dataset and the model was re-trained to compare the precision and recall. The detection result images and visualization heatmap are also shown. Example raw sample images are shown in (**B**), and example processed sample images are shown in (**A**), (**C**), (**D**). Each scale bar represents 10$\mu m$. Each error bar represents the standard deviation of repeatedly training the model on the same dataset by three times. **E**, **F** The decrease in precision and recall caused by each factor was ranked. Removing raw sample images from the training dataset caused the largest drop in model precision, whereas removing 20× images caused the largest drop in model recall. (*$p < 0.1$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$)

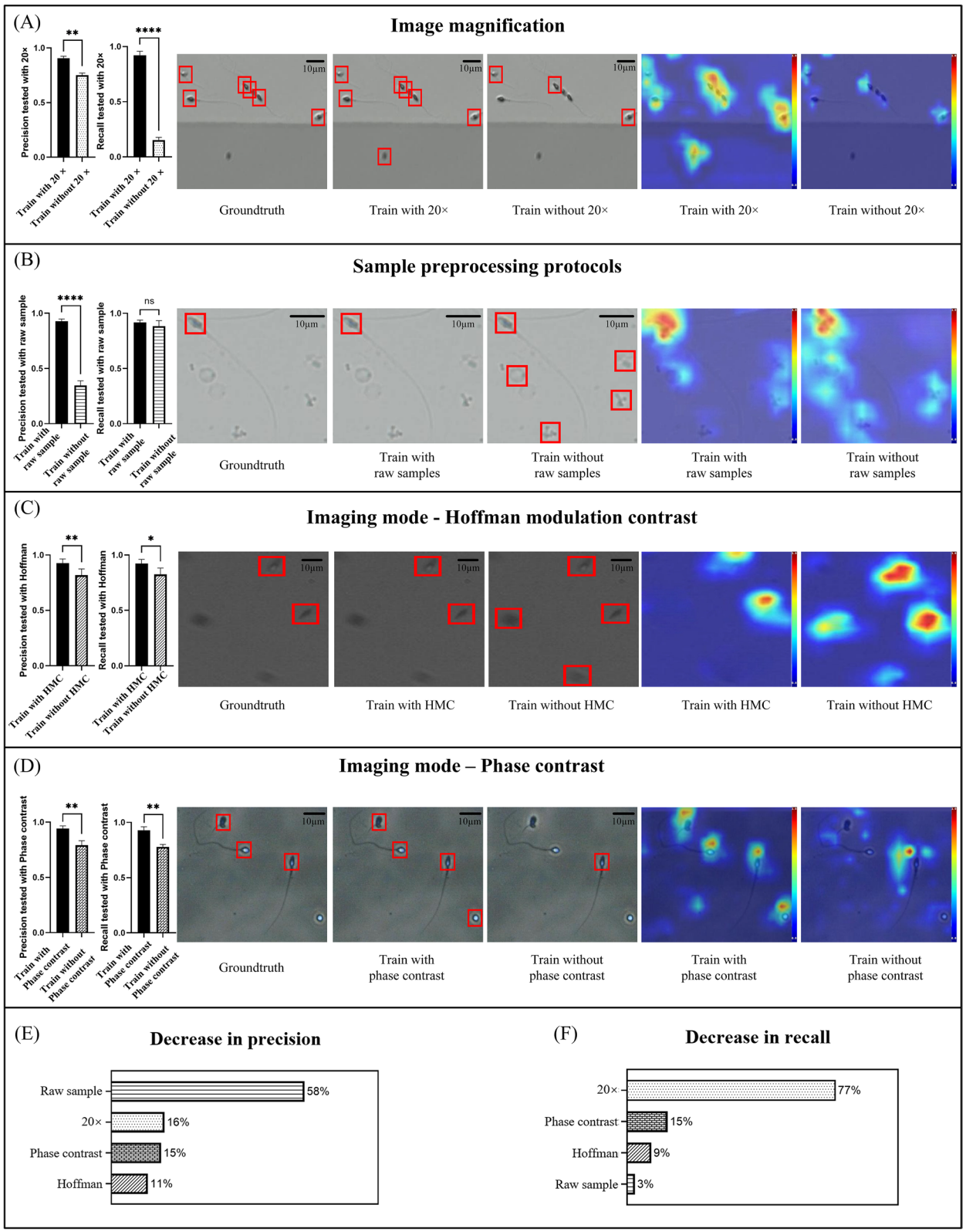


**Fig. 2** (See legend on previous page.)

model focused on similar regions in the weight heatmaps (Fig. 2C, D).

Collectively, among all the factors, removing raw sample images caused the largest drop (58.11%, Fig. 2E) in model precision (the most false-positive detections), while removing 20× images caused the largest drop (76.81%, Fig. 2F) in model recall (the most missed detections). Removing a set of data from the training dataset reduced data richness and resulted in a decrease in both model precision and recall, confirming that richness of data in the training dataset significantly impacts model performance.

#### Improving model generalizability by increasing data richness of the training dataset

Based on the ablation study, we hypothesized that increasing richness of training data would make model performance generalizable under different imaging conditions. Here data richness is twofold: 1) the training dataset should be diverse and include as many features as possible - for a model to correctly detect sperm under different imaging conditions, the model should have seen and learned such features during training to ensure a generalizable model performance; 2) the balance of foreground and background objects in the training dataset should be ensured - the lack of background objects (e.g., non-sperm impurities) decreases model precision.

To test the hypothesis, we included sperm images captured under different imaging magnifications, sample preprocessing protocols, and imaging modes into the training dataset (Supplementary Table 2). The detection model was re-trained (Fig. 4 and Supplementary Fig. 1) and its generalizability was then tested in both internal blind tests on unseen samples and external multicenter validation.

#### Testing the hypothesis via internal blind test of repeated measurement on unseen samples

We first tested the hypothesis by repeatedly detecting sperm from the same sample, but under different imaging and sample preprocessing conditions. The comparison experiments were repeated on 5 raw samples and 5 processed samples. None of these samples were included in the training dataset. Generalizability was evaluated by ICC.

As summarized in Table 2, model precision and recall were both consistently around 91%, regardless of imaging magnification, imaging mode, and raw or process samples. The precision and recall values were also consistent with model training (Supplementary Fig. 1). The maximum standard deviation was 1.66% for precision and 1.77% for recall. In addition, no significant differences were observed in model precision and recall among different imaging magnifications, imaging modes or between raw samples versus process samples ($p > 0.05$). Collectively, by incorporating different imaging and sample preprocessing conditions into a rich training dataset, the model achieved an ICC of 0.97 (95% CI: 0.94-0.99) for precision, and an ICC of 0.97 (95% CI: 0.93-0.99) for recall.

**Table 2** Model performance under repeated measurements with different image acquisition conditions

| Conditions | | Raw sample | | Processed sample | |
|---|---|---|---|---|---|
| | | Precision (%) | Recall (%) | Precision (%) | Recall (%) |
| **Bright field** | 20× | 91.82±0.31 | 90.78±0.43 | 91.73±0.33 | 90.81±1.53 |
| | 40× | 91.77±0.85 | 90.58±0.74 | 91.59±1.56 | 90.57±1.46 |
| **Phase contrast** | 20× | 91.71±0.56 | 90.70±0.34 | 91.84±0.84 | 90.46±0.66 |
| | 40× | 91.73±1.50 | 91.00±1.24 | 91.53±1.66 | 90.73±1.01 |
| **HMC** | 20× | 91.91±0.52 | 90.60±0.25 | 91.53±0.98 | 90.54±1.77 |
| | 40× | 91.63±1.62 | 90.50±1.47 | 91.76±1.21 | 90.44±1.02 |

#### Testing the hypothesis via external validation among three clinics

We further performed an external multicenter clinical validation study to test model generalizability in clinical setups. The pre-trained sperm detection model was tested in three clinics, and in each clinic the model was evaluated in two clinical applications. 1) Raw semen analysis: the model was applied to detect sperm in raw semen samples. This application aids the computation of sperm concentration, which is for computer-aided sperm analysis (CASA) and the diagnosis of male infertility. 2) Processed sample analysis: the model was applied to detect sperm in processed and washed samples. This application is for calculating the dilution ratio necessary for conventional IVF treatments. In each clinic, 10 samples (including 5 raw samples and 5 processed samples) were tested, totaling 30 samples across all sites. This experimental design ensured an evaluation of the model's performance under different sample conditions and clinical setups. The imaging setup in each clinic is summarized in Supplementary Table 3.

Detecting sperm in raw semen is challenging because of the interference of non-sperm cells in semen such as leukocytes and epithelial cells. Similar size and shape could make the algorithm incorrectly identify the sperm cells, leading to a decrease in precision, which may have an impact on sperm concentration calculation. Nonetheless, the model's detection precision of raw samples ranged from 91.40% to 91.78% in the three clinics, and no significant differences were observed among clinics

($p > 0.05$, Fig. 3). A similar result was obtained for model recall (ranged from 89.82% to 90.16%, $p > 0.05$, Fig. 3).

Not surprisingly, for processed samples which had a cleaner background and less interference than raw samples, the model consistently achieved a precision ranged from 91.52% to 91.70% in the three clinics, with no significant differences among clinics ($p > 0.05$, Fig. 3). Model recall for processed samples ranged from 89.98% to 90.16% ($p > 0.05$). Compared with the precision and recall validated during model training, the difference in the three clinics was in the range of 0.02% to 0.20% for precision and −0.32% to −0.14% for recall, and no significant differences were observed ($p > 0.05$, Fig. 3). Collectively, within each clinic, there was no significant difference between the precision or recall tested on raw samples and the processed samples ($p > 0.05$, Fig. 3).

## Discussion

Sperm detection in andrology labs and IVF labs has high reproducibility requirements. Although deep learning models have been developed to automate this tedious task [54], model generalizability remains poorly understood [55]. In clinical research, this type of generalizability is also defined as conceptual reproducibility [56–59] in the literature, referring to the model's ability to generalize and yield consistent outcomes when validating results on novel data from different sources or under various conditions. As deep learning models are increasingly applied in various clinical applications, the generalizability of such models must be investigated before they can be deployed for clinical use. Using sperm detection as a pilot study, this work 1) investigated potential factors affecting generalizability of the deep learning model, and 2) hypothesized strategies for improving the generalizability of object detection models and tested the hypothesis in multiple clinics.

For the first aim, considering deep learning is a data-driven approach and the model learns features from the provided training dataset, we investigated how the training dataset affects model generalizability. In the ablation experiments, the model was re-trained using the dataset ablating/without 20× images. When tested with 20× images as input, the re-trained model showed a significant drop in recall. The drop in recall was also observed when ablating images of raw semen and ablating images captured under the Hoffman modulation contrast and phase contrast imaging mode. These results suggest that richness of the training dataset is necessary for the model's performance to be generalizable under different clinical setups. In other words, for the model to correctly identify an image feature during clinical deployment, the model must have seen and learned such features in the training dataset.

Interestingly, in the ablation study, we noticed that among the three factors, imaging magnification caused the largest drop in recall, with imaging mode ranked next, whereas differences in sample preprocessing protocols did not cause a significant drop (Fig. 2). One potential reason is that the appearance of sperm under 20× vs. 40× was more different than that under Hoffman modulation contrast/phase contrast vs. bright field imaging. Changing magnifications changed the number of pixels occupied by a sperm, and fewer features were available under a smaller magnification. Compared to

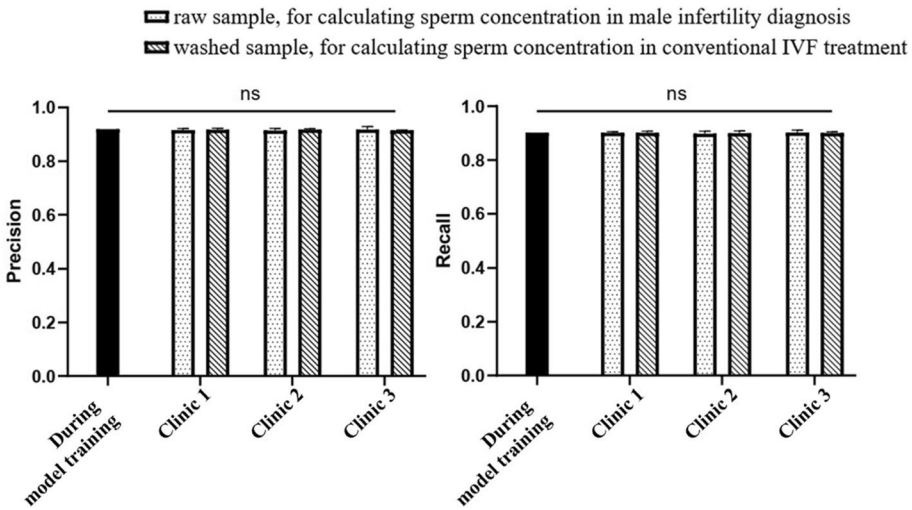


**Fig. 3** Testing the hypothesis in three clinics. The model precision and recall were tested using both raw samples and processed samples in two clinical applications. There was no significant difference in model precision and recall among three clinics as compared to the performance tested in during model training. (ns: not significant, $p > 0.05$)

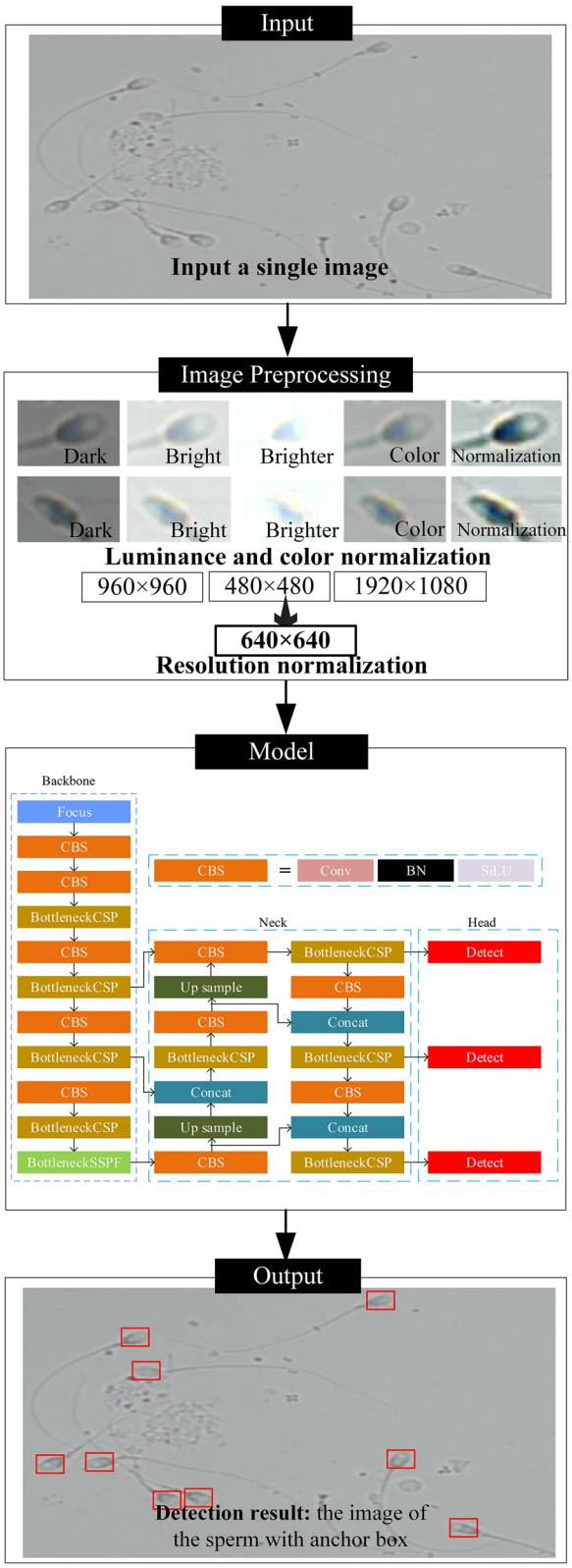


**Fig. 4** Architecture of the deep learning-based sperm detection model. The model takes a single microscopic image as input (raw sample in this example), then uses image preprocessing to normalize image luminance, resolution, and color. Sperm is detected using Yolo v5, one of the state-of-the-art convolutional neural networks for object detection. The model outputs the image of the detected with anchor box markers (bounding boxes) and coordinates of each sperm

magnification-caused changes, Hoffman modulation contrast imaging mainly changed imaging contrast and the resulting images had similar appearance to bright field images. Hence, although the targets to be detected belong to the same class of sperm, the intra-class distance [60, 61] was small for sperm images under different imaging modes and large for different magnifications. Identifying objects with a larger intra-class distance typically requires a more comprehensive and richer dataset [62, 63]. In contrast, the impurities in raw samples did not change the appearance of sperm itself, thus not causing missed detection (recall).

Another aspect of data richness is the richness of positive samples (i.e., sperm) and negative samples (i.e., background, non-sperm cells) in the training dataset. Removing the images of raw semen resulted in the largest drop in model precision. This suggests that balance of positive and negative samples should be ensured in the dataset. In the ablation experiments, the lack of negative samples such as impurities from raw semen resulted in a significantly lower precision when interferences were present. A balanced proportion of positive and negative samples can improve the anti-interference ability of the model, reduce false identification, and improve model generalizability under interference [64, 65].

In addition to the richness of data in the training dataset, the normalization steps during image preprocessing in the model may also contribute to model generalizability. In clinical practice, inconsistencies in the camera and image acquisition schemes lead to different brightness, color (white balance) and resolution of the acquired images. By performing image preprocessing, the brightness and color of the images can be normalized, and the resolution can be resized to the same for inputting into the model (Fig. 4), and the effect of inconsistencies in image acquisition hardware on model performance could be minimized.

For the second aim, according to the hypothesis, we re-trained the model with rich data and tested its generalizability among three clinics. It is worth noting that the objective of this work is not to create a novel model for sperm detection with improved accuracy; instead, we focused on testing the generalizability of state-of-the-art learning models under different clinical setups.

The major difference between this work and existing studies is that in addition to validating model on the retrospectively collected dataset, we further performed prospective experiments to quantify model ICC, and prospective testing among multiple clinics. In existing studies, as a routine for model development and validation, a retrospectively collected dataset is usually split into training, validation, and test sub-datasets. After each step/epoch of model training, the validation sub-dataset is fed into the model to evaluate its accuracy and precision. Hence, existing studies reported the accuracy or precision as the evaluation metric for the developed model. Although such datasets may involve data from multiple clinics, the validation and test sub-datasets were collected under the same conditions as the training sub-dataset. The lack of external validation did not allow the investigation of reproducibility metrics such as ICC.

In addition to the routine model development and validation on the sub-datasets, this work further measured model ICC by repeatedly testing the model on the same sperm samples but imaged under different image acquisition and sample processing conditions. The model achieved an ICC higher than 0.9. In further prospective multicenter validation, although each clinic used different setups, the model consistently achieved a precision and recall higher than 90%, under different image acquisition conditions (magnifications, imaging modes, camera resolution etc.) and different sample processing procedures (raw samples and processed samples).

Our results highlight the importance of considering the imaging conditions used during model development and training. As an explorative study, we aimed to comprehensively include imaging conditions to provide a complete picture of the model's performance across various imaging settings. In practice, clinics are likely to maintain consistent imaging conditions for a given application to ensure standardization and comparability of results. When deploying deep learning models in a single clinic, it is crucial to ensure that the training data closely matches the intended use case. If a model is to be applied across multiple clinics or imaging setups, it is necessary to include a diverse range of imaging conditions in the training data to improve model generalizability.

The approach for testing a model's generalizability from this study paves the foundation for generalizability evaluation of deep learning models in wider andrology and reproductive medicine applications. Our results also draw the attention to the training dataset of deep learning models and suggest that the richness of the training dataset directly impacts the quality of a model.

## Materials and methods

### Sample processing and dataset collection

All human semen samples were collected, processed and tested under the guidance of the World Health Organization protocol, with the approval of the ethics committee (CUHKSZ and three IVF clinics, with IRB numbers listed in section "Testing generalizability among clinics" below) and informed consent of all patients under test. Semen samples were liquefied at room temperature for 30-60 min. Raw samples were untreated, processed samples were purified by the swim-up method, and diluted to a density of 15-200× $10^6$ cells/ml density for analysis to facilitate normal medical tests. All experiments were completed within 3 hours after sperm collection.

For model training in the ablation study and hypothesis testing, a dataset containing images of 7,353 sperm from 60 semen samples was collected using a standard inverted microscope (Nikon ECLIPSE Ti2-E, Nikon Inc.) equipped with a camera (Basler MED ace 2.3, Basler Inc.). The 60 semen samples consist of 35 samples from volunteers and randomly selected medical examiners and 25 samples from infertile patients, all randomly selected, whose semen analysis parameters are summarized in Supplementary Table 1. Three embryologists annotated the sperm images and obtained the location information (i.e., bounding box) of the 7,353 sperm. The collected dataset contained images captured under two different magnifications (20×, 40×) and three imaging modes (bright field, Hoffman modulation contrast and phase contrast). More details of the dataset can be found in Supplementary Table 2.

### Deep learning model for sperm detection

The overall sperm detection model framework is based on YOLO v5, which is one of the state-of-the-art object detection deep learning models (Table 1). The detection model takes a single image as input, and the output is the image of the detected sperm with anchor box markers and coordinates. The neural network structure consists of a backbone module, neck module, and head module, and more details of the network can be found in Fig. 4. The acquired image resolution, luminance, and color may be different in each clinic; hence, an image preprocessing module was added to normalize these factors. The image was resized into 640×640 resolution and fed into the detection model. Similarly, the luminance and color normalization step minimized their impact on model learning.

### Training of the deep learning model

The model was trained based on the dataset containing the 7,353 sperm as mentioned above (part of the dataset for ablation experiments, and the entire dataset for hypothesis testing). During training, in order to avoid overfitting, mosaic data augmentation was used to crop, arrange and stitch images randomly to augment the dataset. In training, the GIoU loss (generalized intersection over union) was used to evaluate the robustness and convergence of the model. The deep learning model was trained using the Pytorch framework (Python 3.9, Pytorch version 1.7.1), on GPU (model: NVIDIA GeForce RTX 3090 24G). The hyperparameters for training were set as follows: the optimizer was Adam, the epochs were 600, the learning rate was 0.001, and the batch size was 64.

### Visualization of model weights

To enhance the interpretability of the model, this study utilized the Gradient Weighted Class Activation Mapping (Grad-CAM) technique [66]. It is a visualization technique for understanding the decision-making process of a deep learning model in an image detection task. Grad-CAM can be integrated with common deep learning frameworks to generate class activation maps by taking a simple image as input, predicting the labels using the full model computation, inserting the global average pooling layer in the model, and computing the gradient of the feature map. The class activation maps generated by Grad-CAM visualize the regions of interest of the model on the input image. For all visualization, Grad-CAM was used in the last Conv layer of the detection model, because the last layer represents the most abstract and decision-relevant features learned by the network.

### Model evaluation

In the study, objective evaluation indicators such as precision, recall, were used to evaluate the performance of the trained sperm detection model. The calculation equations are as follows:

$$\begin{aligned} precision &= \frac{TP}{TP + FP} \\ recall &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \end{aligned} \tag{1}$$

where *TP* is the number of correctly identified sperm targets; *FP* is the number of falsely identify targets; and *FN* is the number of sperm targets that were missed by the model. In the blind test and multicenter validation, at least 200 sperm were detected in each patient sample and benchmarked against manual sperm detection results to calculate *TP*, *FP*, and *FN*.

### Testing generalizability among clinics

Model generalizability was tested among three clinics, including 1) The 3rd Affiliated Hospital of Shenzhen University in Shenzhen, China, with IRB approval number: 2021-LHRMYY-SZLL-012; 2) Reproductive & Genetic Hospital of Citic-Xiangya in Changsha, China, with IRB approval number: LL-SC2021-016; and 3) CReATe Fertility Centre in Toronto, Canada, with IRB approval number: UT35544. It is worth noting that the academic lab (CUHKSZ) for collecting the training dataset was not within these three clinics. Each clinic used a different setup for image acquisition, including different microscopes, cameras, imaging modes and magnifications. A complete list of the setup in each clinic is summarized in Supplementary Table 3.

In each clinic, 5 raw samples and 5 processed samples were processed by lab technicians. For each sample, technicians recorded videos and extracted images from them. Then the model detected the total number of sperm and benchmarked to manual results.

### Statistics

The results were expressed as means and standard deviation. No data points were excluded from the analysis. Statistical analysis was performed with MedCalc 18.3 software (MedCalc Software Ltd.). Differences between the means of two groups were tested with a two-tailed student's t-test, and differences among more than two groups were tested by one-way analysis of variance (ANOVA), followed by Holm-Sidak pairwise comparison for normally distributed data or Dunn's test for non-normally distributed data. Model generalizability in precision and recall was evaluated with ICC (intraclass correlation coefficient). For all tests, $p < 0.05$ (labeled with an asterisk in the figures) was considered as a statistically significant difference.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12958-024-01232-8.

Supplementary Material 1.

**Acknowledgements**
Not applicable.

**Authors' contributions**
Jiaqi Wang: Data collection and analysis, and drafting the manuscript. Yufei Jin: Data analysis and drafting the manuscript. Aojun Jiang: Data collection and analysis. Wenyuan Chen: Acquisition of data. Guanqiao Shan: Acquisition of data. Yifan Gu: Providing clinical guidance and samples, acquisition of data. Yue Ming: Data collection and analysis. Jichang Li: Data collection and analysis. Chunfeng Yue: Testing of algorithms. Zongjie Huang: Testing of algorithms. Clifford Librach: Providing clinical guidance and samples. Ge Lin: Providing clinical guidance and samples. Xibu Wang: Acquisition of data. Huan Zhao:

Providing clinical guidance and samples. Yu Sun: Study design and drafting the manuscript. Zhuoran Zhang: Study design and drafting the manuscript.

**Funding**
This work was supported in part by National Key R &D Program of China (2023YFE0205500), in part by the National Natural Science Foundation of China (62203374), in part by Guangdong Basic and Applied Basic Research Foundation (2021A1515110023), in part by Shenzhen Science and Technology Program (RCBS20210706092254072), and in part by The Chinese University of Hong Kong, Shenzhen (UDF01002141), all to Z. Zhang.

**Availability of data and materials**
Data is provided within the manuscript.

## Declarations

**Ethics approval and consent to participate**
Our study strictly adheres to the ethical principles for medical research involving human subjects, as outlined in the Declaration of Helsinki by the World Medical Association. The study protocol has been reviewed and approved by the respective ethics committees of the participating clinics. This study was tested among three clinics, including 1) The 3rd Affiliated Hospital of Shenzhen University in Shenzhen, China, with IRB approval number: 2021-LHRMYY-SZLL-012; 2) Reproductive & Genetic Hospital of Citic-Xiangya in Changsha, China, with IRB approval number: LL-SC2021-016; and 3) CReATe Fertility Centre in Toronto, Canada, with IRB approval number: UT35544. We have taken all necessary measures to protect the rights and privacy of the participants. It is worth noting that this study is not a clinical trial because the study only involves the use of human sperm samples, without any intervention to patients or evaluating any outcomes on human health. With patients' consents, this study acquired and analyzed images of human sperm samples. The images were then used for evaluating the generalizability of deep learning models for sperm detection.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. [2]Department of Mechanical Engineering, University of Toronto, Toronto, Canada. [3]Institute of Reproductive and Stem Cell Engineering, School of Basic Medical Science, Central South University, Changsha, China. [4]Reproductive & Genetic Hospital of Citic-Xiangya, Changsha, China. [5]School of Medicine, The Chinese University of Hong Kong, Shenzhen, China. [6]Suzhou Boundless Medical Technology Ltd., Co., Suzhou, China. [7]CReATe Fertility Centre, Toronto, Canada. [8]The 3rd Affiliated Hospital of Shenzhen University, Shenzhen, China. [9]Department of Computer Science, University of Toronto, Toronto, Canada. [10]Institute of Biomedical Engineering, University of Toronto, Toronto, Canada. [11]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada.

Received: 3 March 2024 Accepted: 14 May 2024
Published online: 22 May 2024

**References**

1. Gadadhar S, Alvarez Viar G, Hansen JN, Gong A, Kostarev A, Ialy-Radio C, et al. Tubulin glycylation controls axonemal dynein activity, flagellar beat, and male fertility. Science. 2021;371(6525):eabd4914. https://www.science.org/doi/abs/10.1126/science.abd4914.
2. Li X, Li C, Rahaman MM, Sun H, Li X, Wu J, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. Artif Intell Rev. 2022;55(6):4809–78. https://link.springer.com/article/10.1007/s10462-021-10121-0.
3. Marino JL, Moore VM, Rumbold AR, Davies MJ. Fertility treatments and the young women who use them: an Australian cohort study. Hum Reprod. 2011;26(2):473–9. https://academic.oup.com/humrep/article-abstract/26/2/473/593755.
4. Stouffs K, Tournaye H, Van der Elst J, Liebaers I, Lissens W. Is there a role for the nuclear export factor 2 gene in male infertility? Fertil Steril. 2008;90(5):1787–91. https://www.sciencedirect.com/science/article/pii/S001502820703467X.
5. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 2020;21(2):222–32. https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30738-7/fulltext?13570.
6. Kriegeskorte N, Golan T. Neural network models and deep learning. Curr Biol. 2019;29(7):R231–6. https://www.cell.com/current-biology/pdf/S0960-9822(19)30204-0.pdf.
7. Hariton E, Pavlovic Z, Fanton M, Jiang VS. Applications of Artificial Intelligence in Ovarian Stimulation: A Tool for Improving Efficiency and Outcomes. Fertil Steril. 2023. https://www.sciencedirect.com/science/article/abs/pii/S0015028223005198. Accessed 19 May 2024.
8. Fanton M, Nutting V, Solano F, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for predicting the optimal day of trigger during ovarian stimulation. Fertil Steril. 2022;118(1):101–8. https://www.sciencedirect.com/science/article/pii/S0015028222002448.
9. Chandra S, Gourisaria MK, Gm H, Konar D, Gao X, Wang T, et al. Prolificacy Assessment of Spermatozoan via state-of-the-art Deep Learning Frameworks. IEEE Access. 2022;10:13715–27. https://ieeexplore.ieee.org/abstract/document/9693937/.
10. Spencer L, Fernando J, Akbaridoust F, Ackermann K, Nosrati R. Ensembled Deep Learning for the Classification of Human Sperm Head Morphology. Adv Intell Syst. 2022;4(10):2200111. https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202200111.
11. Riordon J, McCallum C, Sinton D. Deep learning for the classification of human sperm. Adv Intell Syst. 2019;111:103342. https://www.sciencedirect.com/science/article/pii/S0010482519302112.
12. Dobrovolny M, Benes J, Langer J, Krejcar O, Selamat A. Study on Sperm-Cell Detection Using YOLOv5 Architecture with Labaled Dataset. Genes. 2023;14(2):451. https://www.mdpi.com/2073-4425/14/2/451.
13. Zhu R, Cui Y, Huang J, Hou E, Zhao J, Zhou Z, et al. YOLOv5s-SA: Light-Weighted and Improved YOLOv5s for Sperm Detection. Diagnostics. 2023;13(6):1100. https://www.mdpi.com/2075-4418/13/6/1100.
14. Zhang Z, Qi B, Ou S, Shi C. Real-Time Sperm Detection Using Lightweight YOLOv5. In: Proceedings of the 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Sichuan, China, December 9-12, 2022. IEEE; 2022. p. 1829–1834. https://ieeexplore.ieee.org/abstract/document/10065602/.
15. Kahveci B, Önen S, Akal F, Korkusuz P. Detection of spermatogonial stem/progenitor cells in prepubertal mouse testis with deep learning. J Assist Reprod Genet. 2023;40(5):1187–95. https://link.springer.com/article/10.1007/s10815-023-02784-1.
16. Lee R, Witherspoon L, Robinson M, Lee JH, Duffy SP, Flannigan R, et al. Automated rare sperm identification from low-magnification microscopy images of dissociated microsurgical testicular sperm extraction samples using deep learning. Fertil Steril. 2022;118(1):90–9. https://www.sciencedirect.com/science/article/pii/S0015028222001959.
17. Targosz A, Myszor D, Mrugacz G. Human oocytes image classification method based on deep neural networks. Biomed Eng OnLine. 2023;22(1):92. https://link.springer.com/article/10.1186/s12938-023-01153-4.
18. Wu C, Yan W, Li H, Li J, Wang H, Chang S, et al. A classification system of day 3 human embryos using deep learning. Biomed Signal Process Control. 2021;70:102943. https://www.sciencedirect.com/science/article/pii/S1746809421005401.
19. Amitai T, Kan-Tor Y, Or Y, Shoham Z, Shofaro Y, Richter D, et al. Embryo classification beyond pregnancy: Early prediction of first trimester miscarriage using machine learning. J Assist Reprod Genet. 2023;40(2):309–22. https://link.springer.com/article/10.1007/s10815-022-02619-5.
20. Bormann CL, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, Kandula H, et al. Performance of a deep learning based neural network in the selection of human blastocysts for implantation. Elife. 2020;9:e55301. https://elifesciences.org/articles/55301.

21. Wan S, Zhao X, Niu Z, Dong L, Wu Y, Gu S, et al. Influence of ambient air pollution on successful pregnancy with frozen embryo transfer: A machine learning prediction model. Ecotoxicol Environ Saf. 2022;236:113444. https://www.sciencedirect.com/science/article/pii/S0147651322002846.
22. Mehrjerd A, Rezaei H, Eslami S, Ratna MB, Khadem Ghaebi N. Internal validation and comparison of predictive models to determine success rate of infertility treatments: a retrospective study of 2485 cycles. Sci Rep. 2022;12(1):7216. https://www.nature.com/articles/s41598-022-10902-9.
23. Blank C, Wildeboer RR, DeCroo I, Tilleman K, Weyers B, De Sutter P, et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. Fertil Steril. 2019;111(2):318–26. https://www.sciencedirect.com/science/article/pii/S0015028218321563.
24. Rienzi L, Cimadomo D, Delgado A, Minasi MG, Fabozzi G, Del Gallego R, et al. Time of morulation and trophectoderm quality are predictors of a live birth after euploid blastocyst transfer: a multicenter study. Fertil Steril. 2019;112(6):1080–1093. e1. https://www.sciencedirect.com/science/article/pii/S0015028219319302.
25. Lee LH, Bradburn E, Craik R, Yaqub M, Norris SA, Ismail LC, et al. Machine learning for accurate estimation of fetal gestational age based on ultrasound images. NPJ Digit Med. 2023;6(1):36. https://www.nature.com/articles/s41746-023-00774-2.
26. Makarious MB, Leonard HL, Vitale D, Iwaki H, Sargent L, Dadu A, et al. Multi-modality machine learning predicting Parkinson's disease. NPJ Park Dis. 2022;8(1):35. https://www.nature.com/articles/s41531-022-00288-w.
27. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med. 2020;3(1):23. https://www.nature.com/articles/s41746-020-0232-8.
28. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. NPJ Digit Med. 2018;1(1):6. https://www.nature.com/articles/s41746-017-0013-1.
29. Zhou J, Hu B, Feng W, Zhang Z, Fu X, Shao H, et al. An ensemble deep learning model for risk stratification of invasive lung adenocarcinoma using thin-slice CT. NPJ Digit Med. 2023;6(1):119. https://www.nature.com/articles/s41746-023-00866-z.
30. Deng Y, Lu L, Aponte L, Angelidi AM, Novak V, Karniadakis GE, et al. Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. NPJ Digit Med. 2021;4(1):109. https://www.nature.com/articles/s41746-021-00480-x.
31. Xu Q, Zhan X, Zhou Z, Li Y, Xie P, Zhang S, et al. AI-based analysis of CT images for rapid triage of COVID-19 patients. NPJ Digit Med. 2021;4(1):75. https://www.nature.com/articles/s41746-021-00446-z.
32. Madani A, Ong JR, Tibrewal A, Mofrad MR. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ Digit Med. 2018;1(1):59. https://www.nature.com/articles/s41746-018-0065-x.
33. Firuzinia S, Afzali SM, Ghasemian F, Mirroshandel SA. A robust deep learning-based multiclass segmentation method for analyzing human metaphase II oocyte images. Comput Methods Prog Biomed. 2021;201:105946. https://www.sciencedirect.com/science/article/pii/S0169260721000201.
34. Jiang VS, Kartik D, Thirumalaraju P, Kandula H, Kanakasabapathy MK, Souter I, et al. Advancements in the future of automating micromanipulation techniques in the IVF laboratory using deep convolutional neural networks. J Assist Reprod Genet. 2023;40(2):251–7. https://link.springer.com/article/10.1007/s10815-022-02685-9.
35. Goss DM, Vasilescu SA, Vasilescu PA, Cooke S, Kim SH, Sacks GP, et al. AI facilitated sperm detection in azoospermic samples for use in ICSI. medRxiv. 2023. https://www.medrxiv.org/content/10.1101/2023.10.25.23297520v1. Accessed 19 May 2024.
36. Kosela M, Aszyk J, Jarek M, Klimek J, Prokop T. Tracking of Spermatozoa by YOLOv5 Detection and StrongSORT with OSNet Tracker. 2022. https://ceur-ws.org/Vol-3583/paper41.pdf. Accessed 19 May 2024.
37. Yuzkat M, Ilhan HO, Aydin N. Detection of sperm cells by single-stage and two-stage deep object detectors. Biomed Signal Process Control. 2023;83:104630. https://www.sciencedirect.com/science/article/pii/S1746809423000630.
38. Zou S, Li C, Sun H, Xu P, Zhang J, Ma P, et al. TOD-CNN: an effective convolutional neural network for tiny object detection in sperm videos. Comput Biol Med. 2022;146:105543. https://www.sciencedirect.com/science/article/pii/S0010482522003353.
39. Mashaal AA, Eldosoky MA, Mahdy LN, Kadry AE. Automatic healthy sperm head detection using deep learning. Int J Adv Comput Sci Appl. 2022;13(4). https://www.proquest.com/openview/33e6627686425a66077a4b3dd0291196/1?pq-origsite=gscholar&cbl=5444811. Accessed 19 May 2024.
40. Siddiqui M, Haugen TB, Riegler MA, Hammer HL. Detecting Human Embryo Cleavage Stages Using YOLO V5 Object Detection Algorithm. In: Nordic Artificial Intelligence Research and Development: 4th Symposium of the Norwegian AI Society, NAIS 2022, Oslo, Norway, May 31–June 1, 2022, Revised Selected Papers. Springer Nature; 2023. pp. 81. https://library.oapen.org/bitstream/handle/20.500.12657/61287/1/978-3-031-17030-0.pdf#page=89.
41. Patil SN, Wali U, Swamy M, Nagaraj S, Patil N. Deep learning techniques for automatic classification and analysis of human in vitro fertilized (IVF) embryos. J Emerg Technol Innov Res. 2018;5(4):100–6. https://www.researchgate.net/profile/Sujata-Patil-9/publication/334596788_Issue_2_JETIR_ISSN-2349-5162_JETIR1802014_Journal_of_Emerging_Technologies_and_Innovative_Research_JETIR_wwwjetir/links/5d341af1299bf1995b3cf1c0/Issue-2-JETIR-ISSN-2349-5162-JETIR1802014-Journal-of-Emerging-Technologies-and-Innovative-Research-JETIR-wwwjetir.pdf.
42. Raudonis V, Paulauskaite-Taraseviciene A, Sutiene K, Jonaitis D. Towards the automation of early-stage human embryo development detection. Biomed Eng Online. 2019;18(1):1–20. https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-019-0738-y.
43. Dobrovolny M, Benes J, Krejcar O, Selamat A. Sperm-cell Detection Using YOLOv5 Architecture. In: International Work-Conference on Bioinformatics and Biomedical Engineering. Springer; 2022. pp. 319–330. https://link.springer.com/chapter/10.1007/978-3-031-07802-6_27.
44. Aristoteles A, Syarif A, Sutyarso S, Lumbanraja F. Identification of human sperm based on morphology using the you only look once version 4 algorithm. Int J Adv Comput Sci Appl. 2022;13(7):424–31. http://repository.lppm.unila.ac.id/43738/.
45. Sato T, Kishi H, Murakata S, Hayashi Y, Hattori T, Nakazawa S, et al. A new deep-learning model using YOLOv3 to support sperm selection during intracytoplasmic sperm injection procedure. Reprod Med Biol. 2022;21(1):e12454. https://onlinelibrary.wiley.com/doi/abs/10.1002/rmb2.12454.
46. Liu G, Shi H, Zhang H, Zhou Y, Sun Y, Li W, et al. Fast Noninvasive Morphometric Characterization of Free Human Sperms Using Deep Learning. Microsc Microanal. 2022;28(5):1767–79. https://academic.oup.com/mam/article-abstract/28/5/1767/6995548.
47. Dai C, Zhang Z, Jahangiri S, Shan G, Moskovstev S, Librach C, et al. Automated motility and morphology measurement of live spermatozoa. Andrology. 2021;9(4):1205–13. https://onlinelibrary.wiley.com/doi/abs/10.1111/andr.13002.
48. Liu H, Zhang Z, Gu Y, Dai C, Shan G, Song H, et al. Development and evaluation of a live birth prediction model for evaluating human blastocysts from a retrospective study. Elife. 2023;12:e83662. https://elifesciences.org/articles/83662.
49. Menapace W, Lathuilière S, Ricci E. Learning to cluster under domain shift. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. Springer; 2020. pp. 736–52. https://link.springer.com/chapter/10.1007/978-3-030-58604-1_44.
50. Vidit V, Engilberge M, Salzmann M. Clip the gap: A single domain generalization approach for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, June 18-22, 2023. IEEE; 2023. p. 3219–29. http://openaccess.thecvf.com/content/CVPR2023/html/Vidit_CLIP_the_Gap_A_Single_Domain_Generalization_Approach_for_Object_CVPR_2023_paper.html.
51. Sun Y, Chong N, Ochiai H. Feature distribution matching for federated domain generalization. In: Asian Conference on Machine Learning. PMLR; 2023. pp. 942–57. https://proceedings.mlr.press/v189/sun23a.html.
52. Elsahar H, Gallé M. To annotate or not? predicting performance drop under domain shift. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November 3-7, 2019. Assoc Comput Linguist. 2019. p. 2163–73. https://aclanthology.org/D19-1222/.

53. Singhal P, Walambe R, Ramanna S, Kotecha K. Domain adaptation: challenges, methods, datasets, and applications. IEEE Access. 2023;11:6973–7020. https://ieeexplore.ieee.org/abstract/document/10017290/.
54. You JB, McCallum C, Wang Y, Riordon J, Nosrati R, Sinton D. Machine learning for sperm selection. Nat Rev Urol. 2021;18(7):387–403. https://www.nature.com/articles/s41585-021-00465-1.
55. Gibney E. Is AI fuelling a reproducibility crisis in science. Nature. 2022;608(7922):250–1. https://www.nature.com/articles/d41586-022-02035-w.
56. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med. 2021;13(586):eabb1655. https://www.science.org/doi/abs/10.1126/scitranslmed.abb1655.
57. Yang J, Soltan AA, Clifton DA. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit Med. 2022;5(1):69. https://www.nature.com/articles/s41746-022-00614-9.
58. Björndahl L, Barratt CL, Mortimer D, Agarwal A, Aitken RJ, Alvarez JG, et al. Standards in semen examination: publishing reproducible and reliable data based on high-quality methodology. Hum Reprod. 2022;37(11):2497–502. https://academic.oup.com/humrep/article-abstract/37/11/2497/6702083.
59. Leushuis E, Van Der Steeg JW, Steures P, Repping S, Bossuyt PM, Blankenstein MA, et al. Reproducibility and reliability of repeated semen analyses in male partners of subfertile couples. Fertil Steril. 2010;94(7):2631–5. https://www.sciencedirect.com/science/article/pii/S0015028210004589.
60. Chen B, Li Z, Ma Y, Wang N, Bai G. ClraCLoss: Intra-class Distance Loss Makes CNN Robust. In: Proceedings of the 2021 10th International Conference on Computing and Pattern Recognition, Shanghai, China, October 15-17, 2021. Assoc Comput Machinery. 2021. p. 290–5. https://dl.acm.org/doi/abs/10.1145/3497623.3497670.
61. Wang Z, Hu Y, Chia LT. Image-to-class distance metric learning for image classification. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11. Springer; 2010. pp. 706–19. https://link.springer.com/chapter/10.1007/978-3-642-15549-9_51.
62. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):1–48. https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-019-0197-0.pdf.
63. Cui Y, Zhou F, Lin Y, Belongie S. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27-30, 2016. IEEE; 2016. p. 1153–1162. http://openaccess.thecvf.com/content_cvpr_2016/html/Cui_Fine-Grained_Categorization_and_CVPR_2016_paper.html.
64. Saini M, Susan S. Deep transfer with minority data augmentation for imbalanced breast cancer dataset. Appl Soft Comput. 2020;97:106759. https://www.sciencedirect.com/science/article/pii/S1568494620306979.
65. Moreno-Barea FJ, Jerez JM, Franco L. Improving classification accuracy using data augmentation on small data sets. Expert Syst Appl. 2020;161:113696. https://www.sciencedirect.com/science/article/pii/S0957417420305200.
66. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, October 22-29, 2017. IEEE; 2017. p. 618–626. http://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.


## Publisher's Note