Testing the reproducibility 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✉], Zhuoran Zhang[1✉].

[1] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, 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, 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, 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.

[*]These authors contributed equally to this work.
✉Corresponding authors. E-mails: zhangzhuoran@cuhk.edu.cn, yu.sun@utoronto.ca, dahuan4302@163.com

ABSTRACT

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 reproducibility of object detection models. This pilot study took sperm analysis as an example. We performed ablation studies using state-of-the-art models for detecting human sperm, and quantitatively reveal how model precision (false-positive detection) and recall (missed detection) are affected by imaging magnification, imaging mode and sample preprocessing protocols, respectively. The results led to the hypothesis that the richness of image acquisition conditions in a training dataset deterministically affects model reproducibility. To test this hypothesis, we enriched the training dataset with a wide range of imaging conditions. The hypothesis was validated by both internal blind test on new samples to quantify model intraclass correlation coefficient and by external multi-center clinical validation in different imaging conditions and different clinical applications. These findings highlight the importance of diversity in a training dataset for model evaluation and suggest that future deep learning models in andrology and reproductive medicine incorporate comprehensive feature sets for enhanced reproducibility across clinics.

Key Words: Deep learning, Semen analysis, Sperm detection, Reproducibility, Multicenter validation

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[5,6,7,8]. Most of these models take images as input, for instance, to evaluate sperm motility, concentration, and morphology for selecting high-quality sperm for fertilization[9,10,11] or for diagnosing male infertility[12,13,14] , 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[22,23,24].

Despite the potential of deep learning models for advancing clinical practice, existing studies focused on improving model accuracy[25,26,27,28] or precision[29,30,31,32] while little attempt has been made to investigate model reproducibility, 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 reproducibility 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[35,36,37,38,39], and embryo[40,41,42,20]) in images. Different clinics, however, use different image acquisition conditions (e.g., microscope brands and models, imaging modes[43,44,45], magnifications[9,33], illumination intensity, and camera resolutions[13,14,15,39] 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[12,13,14 ,15,35,36,37,38,39,43,44,45,] 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 reproducibility of the developed model under different clinical setups. To address this question, what is needed is prospective validation and testing of model reproducibility. 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.

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

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	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	448×448	40×	[45]
		YOLO	Single center	Hoffman	1664×1664	/	[35]
		YOLO	Single center	/	/	/	[36]
		YOLO	Single center	Bright field	640×640	10×	[37]
		YOLO VGG	Single center	Bright field	698×528	20×	[38]
		VGG	Single center	Bright field	150×150	40×	[39]
		CNN	Single center	DIC	/	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]

RESULTS

Investigating factors that impact model reproducibility

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 reproducibility 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 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. 1A). Model recall also significantly dropped from 92.08% to 15.27% ($p<0.0001$, Fig. 1A). 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. 1A. 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. 1B). 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 images from the training dataset, model precision and recall also dropped (Fig. 1C). 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. 1D). Hoffman and phase contrast imaging

modes mainly changed image contrast, and the resulting images were largely similar to brightfield images. Among the two experiments, the model focused on similar regions in the weight heatmaps (Fig. 1C, 1D).

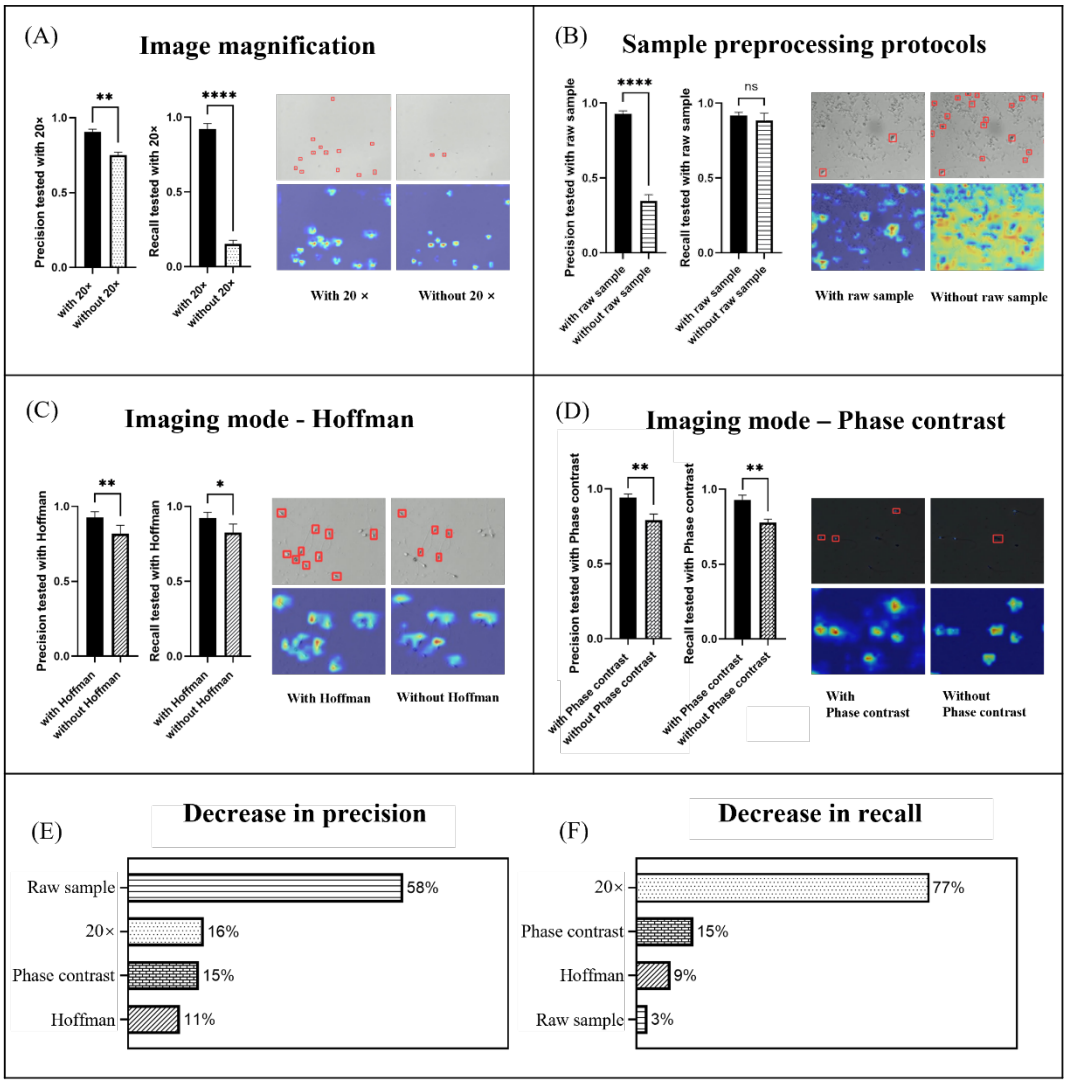

Fig. 1 Ablation studies were performed to investigate how model reproducibility is affected by imaging magnification, imaging mode, and sample preprocessing protocols. (A-D) In the ablation experiment, each investigated factor was removed from the training dataset and the model was re-trained to compare the precision and recall. The detection result images and visualization heatmap are also shown. 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)

Collectively, among all the factors, removing raw sample images caused the largest drop (58.11%, Fig. 1E) in model precision (the most false-positive

detections), while removing 20× images caused the largest drop (76.81%, Fig. 1F) 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 reproducibility 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 reproducible 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 reproducible 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 (Supplementary Fig. 1 and Supplementary Fig. 2) and its reproducibility 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. Reproducibility was evaluated by intraclass correlation coefficient (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. 2). 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
Hoffman	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
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

Testing the hypothesis via external validation among three clinics

We further performed an external multicenter clinical validation study to test model reproducibility in clinical setups. All test data were taken from a random sample of patients attending three clinics, including medical examiners and infertile patients. Each clinic used a different setup for image acquisition (Supplementary Table 3). In each clinic, the sperm detection model was tested under two clinical applications: 1) detecting sperm in raw semen to calculate sperm concentration for computer-aided sperm analysis (CASA) and male infertility diagnosis; and 2) detecting sperm in processed and washed samples to calculate dilution ratio for conventional *in vitro* fertilization treatment. In each clinic, 5 raw samples and 5 processed samples were tested.

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. 2). A similar result was obtained for model recall (ranged from 89.82% to 90.16%, $p>0.05$, Fig. 2).

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. 2). 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. 2). 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. 2).

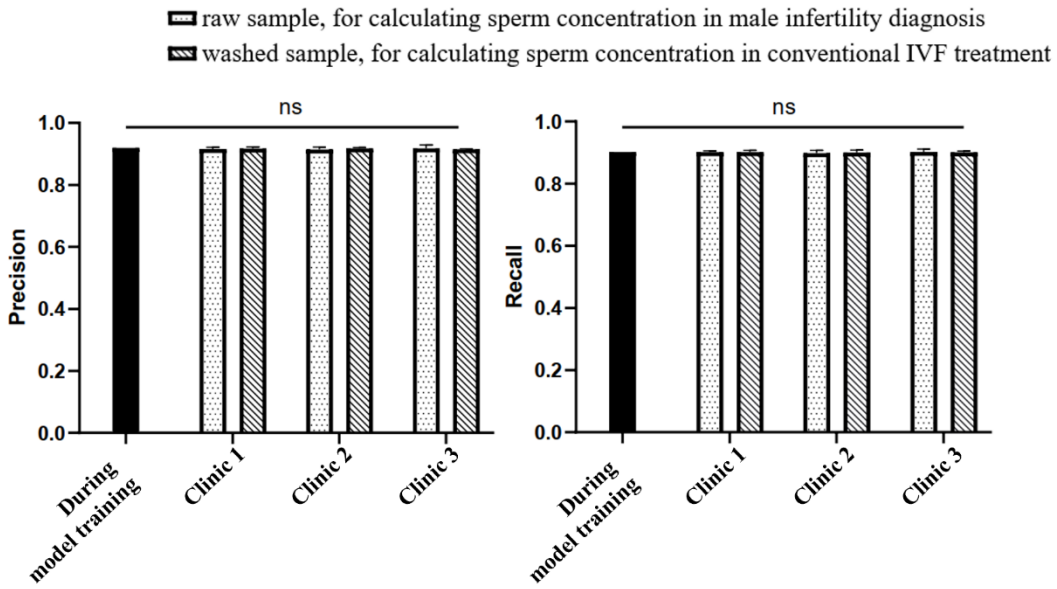

Fig. 2 Testing the hypothesis in three clinics. The model precision and recall were tested using both raw samples and processed samples in two clinical applications. There was no significant difference in model precision and recall among three clinics as compared to the performance tested in during model training. (ns: not significant, p>0.05)

DISCUSSION

Sperm detection in andrology labs and IVF labs has high reproducibility requirements[49,50,51]. Although deep learning models have been developed to automate this tedious task[52], model reproducibility remains poorly understood[53]. As deep learning models are increasingly applied in various clinical applications, the reproducibility 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 reproducibility of the deep learning model, and 2) hypothesized strategies for improving the reproducibility 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 reproducibility. 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 and phase contrast imaging mode. These results suggest that richness of the training dataset is necessary for the model's performance to be reproducible 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. 1). One potential reason is that the appearance of sperm under 20× vs. 40× was more different than that under Hoffman/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 magnification-caused changes, Hoffman 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[54,55] 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[56,57]. 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 reproducibility under interference[58,59.]

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 reproducibility. 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 (Supplementary Fig. 1), 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 reproducibility 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 reproducibility 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).

The approach for testing a model's reproducibility from this study paves the foundation for reproducibility 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 reproducibility 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 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[14], and more details of the network can be found in Supplementary Fig. 1. 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 1200×900 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[60]. 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. In this study, Grad-CAM was used in the last Conv layer of the detection model.

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:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

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 Reproducibility among Clinics

Model reproducibility 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 reproducibility 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.

Data availability

The dataset during the current study is available in the [github] repository and can be accessed via this link [https://github.com/jiaqiwang-rex/Sperm-datasets-for-training].

AUTHOR 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.

ACKNOWLEDGMENT

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.

CONFLICT OF INTEREST STATEMENT

The authors disclose no conflict of interest.

REFERENCES

1. Gadadhar, S. *et al.* Tubulin glycylation controls axonemal dynein activity, flagellar beat, and male fertility. *Science* **371**, eabd4914 (2021).
2. Li, X. *et al.* A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artificial Intelligence Review* **55**, 4809-4878 (2022).
3. Marino, J.L., Moore, V.M., Rumbold, A.R. & Davies, M.J. Fertility treatments and the young women who use them: an Australian cohort study. *Human Reproduction* **26**, 473-479 (2011).
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? *Fertility and sterility* **90**, 1787-1791 (2008).
5. Ström, P. *et al.* Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology* **21**, 222-232 (2020).
6. Kriegeskorte, N. & Golan, T. Neural network models and deep learning. *Current Biology* **29**, R231-R236 (2019).
7. Hariton, E., Pavlovic, Z., Fanton, M. & Jiang, V.S. Applications of Artificial Intelligence in Ovarian Stimulation: A Tool for Improving Efficiency and Outcomes. *Fertility and Sterility* (2023).
8. Fanton, M. *et al.* An interpretable machine learning model for predicting the optimal day of trigger during ovarian stimulation. *Fertility Sterility* **118**, 101-108 (2022).
9. Chandra, S. *et al.* Prolificacy Assessment of Spermatozoan via state-of-the-art Deep Learning Frameworks. *IEEE Access* **10**, 13715-13727 (2022).
10. Spencer, L., Fernando, J., Akbaridoust, F., Ackermann, K. & Nosrati, R. Ensembled Deep Learning for the Classification of Human Sperm Head Morphology. *Advanced Intelligent Systems* **4**, 2200111 (2022).
11. Riordon, J., McCallum, C. & Sinton, D. Deep learning for the classification of human sperm. *Advanced Intelligent Systems* **111**, 103342 (2019).
12. Dobrovolny, M., Benes, J., Langer, J., Krejcar, O. & Selamat, A. Study on Sperm-Cell Detection Using YOLOv5 Architecture with Labaled Dataset. *Genes* **14**, 451 (2023).
13. Zhu, R. *et al.* YOLOv5s-SA: Light-Weighted and Improved YOLOv5s for Sperm Detection. *Diagnostics* **13**, 1100 (2023).
14. Zhang, Z., Qi, B., Ou, S. & Shi, C. in 2022 IEEE 8th International Conference on Computer and Communications (ICCC) 1829-1834 (IEEE, 2022).
15. Kahveci, B., Önen, S., Akal, F. & Korkusuz, P. Detection of spermatogonial stem/progenitor cells in prepubertal mouse testis with deep learning. *Journal of Assisted Reproduction and Genetics* **40**, 1187-1195 (2023).

16. Lee, R. *et al.* Automated rare sperm identification from low-magnification microscopy images of dissociated microsurgical testicular sperm extraction samples using deep learning. *Fertility and Sterility* **118**, 90-99 (2022).
17. Targosz, A., Myszor, D. & Mrugacz, G. Human oocytes image classification method based on deep neural networks. *BioMedical Engineering OnLine* **22**, 92 (2023).
18. Wu, C. *et al.* A classification system of day 3 human embryos using deep learning. *Biomedical Signal Processing and Control* **70**, 102943 (2021).
19. Amitai, T. *et al.* Embryo classification beyond pregnancy: Early prediction of first trimester miscarriage using machine learning. *Journal of Assisted Reproduction and Genetics* **40**, 309-322 (2023).
20. Bormann, C.L. *et al.* Performance of a deep learning based neural network in the selection of human blastocysts for implantation. *Elife* **9**, e55301 (2020).
21. Wan, S. *et al.* Influence of ambient air pollution on successful pregnancy with frozen embryo transfer: A machine learning prediction model. *Ecotoxicology and Environmental Safety* **236**, 113444 (2022).
22. Mehrjerd, A., Rezaei, H., Eslami, S., Ratna, M.B. & Khadem Ghaebi, N. Internal validation and comparison of predictive models to determine success rate of infertility treatments: a retrospective study of 2485 cycles. *Scientific reports* **12**, 7216 (2022).
23. Blank, C. *et al.* Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. *Fertility and sterility* **111**, 318-326 (2019).
24. Rienzi, L. *et al.* Time of morulation and trophectoderm quality are predictors of a live birth after euploid blastocyst transfer: a multicenter study. *Fertility and sterility* **112**, 1080-1093. e1081 (2019).
25. Lee, L.H. *et al.* Machine learning for accurate estimation of fetal gestational age based on ultrasound images. *NPJ digital medicine* **6**, 36 (2023).
26. Makarious, M.B. *et al.* Multi-modality machine learning predicting Parkinson's disease. *NPJ Parkinson's Disease* **8**, 35 (2022).
27. Kiani, A. *et al.* Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine* **3**, 23 (2020).
28. Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine* **1**, 6 (2018).
29. Zhou, J. *et al.* An ensemble deep learning model for risk stratification of invasive lung adenocarcinoma using thin-slice CT. *NPJ digital medicine* **6**, 119 (2023).
30. Deng, Y. *et al.* Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. *NPJ digital medicine* **4**, 109 (2021).

31. Xu, Q. *et al.* AI-based analysis of CT images for rapid triage of COVID-19 patients. *NPJ digital medicine* **4**, 75 (2021).
32. Madani, A., Ong, J.R., Tibrewal, A. & Mofrad, M.R. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ digital medicine* **1**, 59 (2018).
33. Firuzinia, S., Afzali, S.M., Ghasemian, F. & Mirroshandel, S.A. A robust deep learning-based multiclass segmentation method for analyzing human metaphase II oocyte images. *Computer Methods and Programs in Biomedicine* **201**, 105946 (2021).
34. Jiang, V.S. *et al.* Advancements in the future of automating micromanipulation techniques in the IVF laboratory using deep convolutional neural networks. *Journal of Assisted Reproduction and Genetics* **40**, 251-257 (2023).
35. Goss, D.M. *et al.* AI facilitated sperm detection in azoospermic samples for use in ICSI. *medRxiv*, 2023.2010. 2025.23297520 (2023).
36. Kosela, M., Aszyk, J., Jarek, M., Klimek, J. & Prokop, T. Tracking of Spermatozoa by YOLOv5 Detection and StrongSORT with OSNet Tracker. (2022).
37. Yuzkat, M., Ilhan, H.O. & Aydin, N. Detection of sperm cells by single-stage and two-stage deep object detectors. *Biomedical Signal Processing and Control* **83**, 104630 (2023).
38. Zou, S. *et al.* TOD-CNN: An effective convolutional neural network for tiny object detection in sperm videos. *Computers in Biology and Medicine* **146**, 105543 (2022).
39. Mashaal, A.A., Eldosoky, M.A., Mahdy, L.N. & Kadry, A.E. Automatic healthy sperm head detection using deep learning. *International Journal of Advanced Computer Science and Applications* **13** (2022).
40. Siddiqui, M., Haugen, T.B., Riegler, M.A. & Hammer, H.L. 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 81 (Springer Nature, 2023).
41. Patil, S.N., 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* **5**, 100-106 (2018).
42. Raudonis, V., Paulauskaite-Taraseviciene, A., Sutiene, K. & Jonaitis, D. Towards the automation of early-stage human embryo development detection. *Biomedical engineering online* **18**, 1-20 (2019).
43. Dobrovolny, M., Benes, J., Krejcar, O. & Selamat, A. in International Work-Conference on Bioinformatics and Biomedical Engineering 319-330 (Springer, 2022).
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. *International Journal of Advanced Computer Science and Applications* **13**, 424-431 (2022).
45. Sato, T. *et al.* A new deep‐learning model using YOLOv3 to support sperm selection during intracytoplasmic sperm injection procedure. *Reproductive Medicine and Biology* **21**, e12454 (2022).
46. Liu, G. *et al.* Fast Noninvasive Morphometric Characterization of Free Human Sperms Using Deep Learning. *Microscopy and Microanalysis* **28**, 1767-1779 (2022).
47. Dai, C. *et al.* Automated motility and morphology measurement of live spermatozoa. *Andrology* **9**, 1205-1213 (2021).
48. Liu, H. *et al.* Development and evaluation of a live birth prediction model for evaluating human blastocysts from a retrospective study. *Elife* **12**, e83662 (2023).
49. Björndahl, L. *et al.* Standards in semen examination: publishing reproducible and reliable data based on high-quality methodology. *Human Reproduction* **37**, 2497-2502 (2022).
50. Leushuis, E. *et al.* Reproducibility and reliability of repeated semen analyses in male partners of subfertile couples. *Fertility and sterility* **94**, 2631-2635 (2010).
51. McDermott, M.B. *et al.* Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine* **13**, eabb1655 (2021).
52. You, J.B. *et al.* Machine learning for sperm selection. *Nature Reviews Urology* **18**, 387-403 (2021).
53. Gibney, E. Is AI fuelling a reproducibility crisis in science. *Nature* **608**, 250-251 (2022).
54. Chen, B., Li, Z., Ma, Y., Wang, N. & Bai, G. in Proceedings of the 2021 10th International Conference on Computing and Pattern Recognition 290-295 (2021).
55. Wang, Z., Hu, Y. & Chia, L.-T. in Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11 706-719 (Springer, 2010).
56. Shorten, C. & Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *Journal of big data* **6**, 1-48 (2019).
57. Cui, Y., Zhou, F., Lin, Y. & Belongie, S. in Proceedings of the IEEE conference on computer vision and pattern recognition 1153-1162 (2016).
58. Saini, M. & Susan, S. Deep transfer with minority data augmentation for imbalanced breast cancer dataset. *Applied Soft Computing* **97**, 106759 (2020).
59. Moreno-Barea, F.J., Jerez, J.M. & Franco, L. Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications* **161**, 113696 (2020).
60. Selvaraju, R.R. *et al.* in IEEE International Conference on Computer Vision (ICCV) 618-626 (2017).