ORIGINAL ARTICLE

A new deep-learning model using YOLOv3 to support sperm selection during intracytoplasmic sperm injection procedure

Takuma Sato[1] | Hiroshi Kishi[1] | Saori Murakata[1] | Yuki Hayashi[1] |
Toshiyuki Hattori[2] | Shinji Nakazawa[3] | Yusuke Mori[1] | Miwa Hidaka[1] |
Yuta Kasahara[1] | Atsuko Kusuhara[1] | Kayo Hosoya[4] | Hiroshi Hayashi[4] |
Aikou Okamoto[1]

[1]Department of Obstetrics and Gynecology, The Jikei University School of Medicine, Tokyo, Japan

[2]Technology Innovation, GlobalOlympus Corporation, Tokyo, Japan

[3]LPIXEL Inc., Tokyo, Japan

[4]Keiai Reproductive and Endosurgical Clinic, Saitama, Japan

Correspondence
Hiroshi Kishi, Department of Obstetrics and Gynecology, The Jikei University School of Medicine, Japan, Zip Code: 105-8461 3-25-8 Nishi-Shimbashi, Tokyo.
Email: hrskishi@gmail.com

Abstract

Purpose: To create and evaluate a machine-learning model for YOLOv3 that can simultaneously perform morphological evaluation and tracking in a short time, which can be adapted to video data under an inverted microscope.

Methods: Japanese patients who underwent intracytoplasmic sperm injection at the Jikei University School of Medicine and Keiai Reproductive and Endosurgical Clinic from January 2019 to March 2020 were included. An AI model that simultaneously performs morphological assessment and tracking was created and its performance was evaluated.

Results: For morphological assessment, the sensitivity and positive predictive value (PPV) of this model for abnormal sperm were 0.881 and 0.853, respectively. The sensitivity and PPV for normal sperm were 0.794 and 0.689, respectively. For tracking performance, among the 51 objects, 40 (78.4%) were mostly tracked, 11 (21.6%) were partially tracked, and 0 (0%) were mostly lost.

Conclusions: This study showed that evaluating sperm morphology while tracking in a single model is possible by training YOLO v3. This model could acquire time-series data of one sperm, which will assist in acquiring and annotating sperm image data.

KEYWORDS
deep learning, infertility, intracytoplasmic sperm injection, sperm morphology, sperm motility

1 | INTRODUCTION

Infertility is defined as a condition in which a couple who wants a child does not become pregnant for more than 1 year, and about 15% of couples is diagnosed with infertility. Nearly 50% of infertile couples has a component of male factor, and almost 30% is caused solely due to the male factor,[1] and the number of patients with poor sperm analysis is increasing worldwide.[2] The presence of severe male factors necessitates the use of intracytoplasmic sperm injection (ICSI). During ICSI, embryologists evaluate sperm movement and morphology to determine whether to use the sperm for fertilization operations. The World Health Organization Laboratory Manual and Menkveld [3,4] reported the characteristics of normal morphological sperm, and in clinical practice, embryologists have learned these

criteria empirically and select sperm based on their own criteria. Therefore, it is said that the morphological evaluation of sperm in ICSI is personal, subjective, and nonreproducible.[5] Thus, the development of a sperm assessment method using machine learning is expected.

Several machine-learning sperm assessments have been reported for the evaluation of morphology and motility. Studies have evaluated the morphological classification performance of stained sperm images such as SCIAN, HuSHeM, and SMIDS using various machine-learning methods for published databases,[6–9] and research has examined the performance of the morphological classification of unstained sperm image database called MHSMA for each acrosome, head, and vacuole site.[10,11] Studies evaluating sperm motility reported the use of images with a cover glass taken with an upright microscope.[12] When performing ICSI, the embryologist uses an inverted microscope to observe the head and neck of unstained sperm at a low magnification of 400× while at the same time evaluating motility and morphology. To date, there have been no reports of a machine-learning model that can support such a series of work procedures at the same time. In addition, the studies reported so far have been modeled on data sets created at specific orientations and angles suitable for assessing sperm morphology. Models created with such data sets might have a risk of not being able to handle moving sperm.

In this study, we used the residual specimens of ICSI patients to acquire moving image data under an inverted microscope at a low magnification of 400× and created a new data set of unstained sperm images. This data set is intended for practical clinical use of deep-learning models and includes images of orientation, angle, and focus from which it is difficult to evaluate morphology. Based on this data set, we created and verified the performance of a new deep-learning model that performs the morphological classification of sperm and sperm-tracking in a short time.

2 | MATERIALS AND METHODS

2.1 | Data set

We introduce a new data set referred to as the Jikei sperm data set (JSD). This data set consists of high-quality images of 4625 unstained sperm, of Japanese men, evaluated on an RGB scale. Figure S1 shows some examples of the JSD, and Table 1 shows the distribution of the samples in this data set. The method of generating the data set is described in Supporting Information 1 and Figures S2 and S3. The samples in the JSD can be divided into two groups: Groups A and B. Group A contains images that are generally out of focus, contains images showing a large number of sperm in one frame, and includes obvious abnormal sperm such as those with a bent neck. Group B mainly contains images of well-focused abnormal and normal sperms. Group A represents the situation in which specific sperms are selected from a large number of sperms (Figure S1A), and Group

B represents the situation in which sperms after immobilization are selected using a manipulator near the center of the angle of view (Figure S1B).

2.2 | Deep-learning model and learning method

We used YOLO v3, a deep-learning model with the addition of a detection head (Figure 1).[13,14]

To support embryologists in clinical practice, real-time assessment is a priority; thus, we used YOLOv3, which can perform object detection and classification in a short time, in this study. The detection head is highlighted in Figure 1 and is useful for extracting features from high-resolution layers to detect small features such as vacuoles. The implementation of YOLOv3 used AlexeyAB/darknet (https://github.com/AlexeyAB/darknet) to run on Windows.

Artificial intelligence (AI) learning consists of preprocessing, augmentation, and training (Figure S4), and their details are described below.

2.2.1 | Preprocessing

For the input, we used the RGB image of the frame (2448 × 1920 pixels) of the video taken with a microscope, cropped at 960 × 960 pixels centering on the annotated sperm, and resized to 448 × 448 pixels.

2.2.2 | Augmentation

To generate image data to supplement the variation of normal and abnormal sperm, we performed the following augmentation on the 448 × 448 images: Exposure, saturation, and hue were ±10% each; image shift was ±30% both vertically and horizontally; the scale conversion of the input image size was ±20%; and a mix-up was used to combine the images and labels for any two data.[3]

To improve the accuracy of cases in which abnormal sperm was detected due to a wide head, we used the data generated by applying an elastic transform to the normal sperm data. Elastic transform is a transformation that randomly distorts the vertical and horizontal grids and trains data that are largely distorted as abnormal and data that are distorted small as normal (Figure S5).

TABLE 1 Distribution of samples in the proposed data set

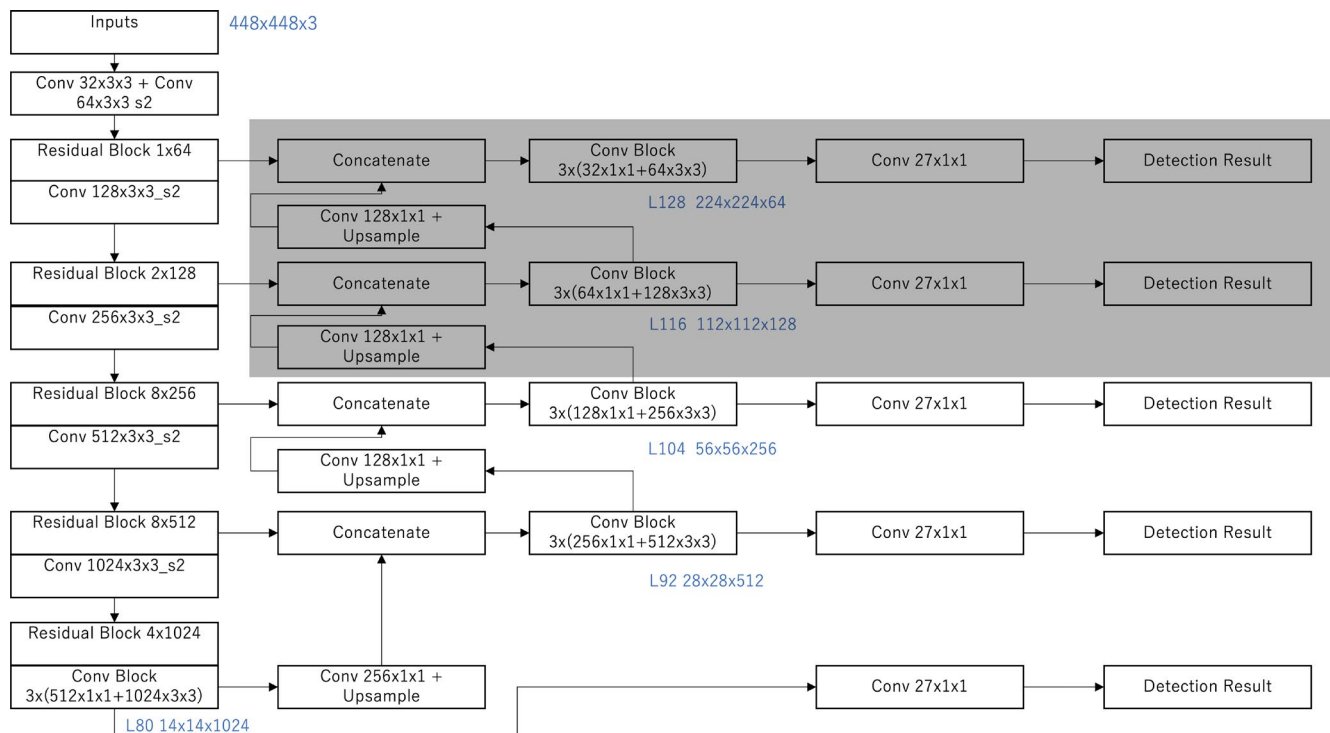| Class	Group A	Group B	Total
Normal	658	270	928
Abnormal	1287	966	2248
Unclassifiable	1095	0	1095
Vacuole	0	354	354

FIGURE 1 Architecture of the model with detection head added to YOLO v3. The shading in the figure represents the detection head. Because this structure enables the extraction of features from a layer with high resolution, it is suitable for detecting small features such as vacuoles

2.2.3 | Training

The image generated by preprocessing and augmentation was input to YOLOv3 as training data. The learning parameters were batch size 64, iteration 10 000, and initial learning rate 0.0010. The loss function was binary cross-entropy loss with focal loss. Focal loss is a method that scales the loss that is easy to recognize relatively small and the loss that is difficult to recognize relatively large and reflects it in learning.[15] A stochastic gradient descent optimizer was used for learning. The parameters of SGD were momentum 0.90 and weight decay 0.005. Furthermore, to converge the learning, we introduced scheduling of the learning rate, and the learning rate was multiplied by 0.10 at 8000 and 9000 iterations, respectively. Of the 1000 iteration training, we adopted the one with the maximum mean average precision of the validation data as the result of the learning.

2.3 | AI prediction

AI prediction consists of prediction, postprocessing, and the prediction result (Figure S6), and their details are described below.

2.3.1 | Prediction

The image data output in real-time from the microscope camera was converted to an image size of 1440 × 1120 pixels and used as an input to the AI. There are four reasons to convert to 1440 × 1120 pixels: it is a multiple of 32, which is in accordance with the restrictions of YOLOv3; it is a conversion close to the scale conversion (448/960 = 0.467) performed at the time of learning; it maintains the aspect ratio of the image (2448/1920 = 1.275, 1440/1,120 = 1.286); and it meets the constraints of the GPU memory used (<10 GB).

2.3.2 | Postprocessing

Because YOLOv3 might detect one object in duplicate, we performed nonmaximum suppression to integrate objects with an intersection over the union of ≥50%.

2.3.3 | Prediction result

The detection thresholds were fixed at 0.25 for normal, abnormal, and unclassifiable and 0.05 for vacuole. The detection thresholds are the thresholds at which a label cannot detect sperm if the prediction probability is less than the threshold value. For the classification of normal, abnormal, and unclassifiable, we used the label with the highest prediction probability as the final prediction result. We lowered the threshold in the vacuole because we assumed that the detection rate would be increased, and the final decision would be left to the embryologist. In addition, the time required for the judgment display was calculated.

2.4 | Evaluation method of AI performance of morphological assessment

We performed fivefold cross-validation to evaluate the proposed model. In fivefold cross-validation, the annotated image data were randomly divided into five groups. Image data extracted from the same video were divided into the same group. We used 9/10 of the four groups as training, 1/10 as validation, and one group as the test. The model created by training data was applied to the test data, and the performance of the model was evaluated. As a result, five models were generated, and each was evaluated. The mean sensitivity and positive predictive value (PPV) of the five models were used as evaluation indexes. The "unclassifiable" label was not included in the evaluation data of the model because this label is a classification set for tracking sperm that cannot be judged as normal or abnormal, and its discrimination performance is not of interest to this study. The definition of "sensitivity" and "PPV" is given in Supporting Information 2 and Table S1.

The PPV was used as an important index for minimizing the negative impact on the embryologist by making a prediction that AI was different from the "ground truth." For vacuole detection, sensitivity, PPV, and specificity were calculated in the same manner.

2.5 | Method for visual explanation

We used gradient-weighted class activation mapping (Grad-CAM) to evaluate this model.[16] This is a heat map showing the areas of interest when the model makes classification predictions. In this study, Grad-CAM was applied to two layers (L92 and L104) with different resolutions (Figure 1). L104 is a layer that detects fine features compared with L92, and in vacuole detection, it is ideal to show that the Grad-CAM in L104 is more focused on the vacuole region.

2.6 | Method for tracking sperm

YOLO v3 performed object detection in each frame of the video, and object tracking was performed by associating the same object's bounding box between frames. SORT was used as the tracking method.[17] A description of SORT is provided in Supporting Information 3.

2.7 | Evaluation method of tracking performance

For the tracking evaluation, one video was used. For the model evaluating the tracking performance, a fold1 model that does not use this video as training data was used. There were 51 sperm in the video, and we evaluated what percentage of all 152 frames could be tracked. Those detected in 80% or more frames were classified as mostly tracked, those detected in 20% to 80% of frames were

TABLE 2 Average performance of fivefold cross-validation

	Abnormal	Normal	Vacuole
Sensitivity	0.881	0.794	0.537
PPV	0.853	0.689	0.585

Abbreviation: PPV—positive predictive value.

classified as partially tracked, and those detected in less than 20% of frames were classified as mostly lost. We also measured the ID switch, the number of times the tracking was interrupted, and the sperm was recognized as another sperm. In addition, false positives, which recognize the place where there is no sperm as sperm, and false negatives, which do not recognize the place where sperm is present as sperm, were also evaluated in all frames. We also calculated the multiple object tracking precision (MOTP) and the multiple object tracking accuracy (MOTA),[18] which are often used to evaluate tracking performance. The definitions of "MOTP" and "MOTA" are given in Supporting Information 4.

3 | RESULTS

3.1 | Performance of sperm morphological assessment

Table 2 displays the average sensitivity and PPV of the fivefolds, and the sum of the confusion matrices in the proposed model is presented in Table 3. Table S2 shows the performance and confusion matrix of each fold.

The sensitivity of this model to abnormal sperm was 0.881, and the PPV was 0.853, which were extremely high values. (Video S1) The detection of abnormalities indicates that about 90% of the cases can be judged as being equivalent to those of the embryologist. It was clarified that by using a data set with abundant variations, it is possible for the model to learn to an accuracy that is sufficient for use in clinical practice. The sensitivity to normal sperm was 0.794, and the PPV was 0.689, which is slightly lower than those detected by abnormal sperm; however, the accuracy was relatively high. (Video S2) In this model, normal sperm candidates can be predicted simultaneously among multiple motile sperms, and the embryologist made the final judgment, using a manipulator to determine whether to select the sperm for ICSI.

The sensitivity of vacuole detection was 0.537, and the PPV was 0.585, which was not very high. Even if vacuoles are detected, their use in ICSI depends on their size and position. Thus, the threshold is set to 0.05 with priority given to detection sensitivity. As a result, the PPV was lower than in other predictions, and the specificity was as high as 0.876. It is considered that the detection itself is difficult because of the small number of pixels of the vacuole, and in cases of multiple vacuoles, the sensitivity and PPV are lowered because of the inability to correctly recognize the second and subsequent vacuoles. The inference time required for the proposed model was 0.246 s. This means that it is possible to obtain a judgment with a

TABLE 3 Sum of confusion matrixes of proposed model for evaluation on the test set

Label	Actual class Abnormal	Normal	Unclassifiable	Vacuole	No object
Abnormal	647	51	0	0	50
Normal	77	215	0	0	16
Unclassifiable	0	4	0	0	2
Vacuole	1	0	0	154	95
No object	0	0	0	125	0

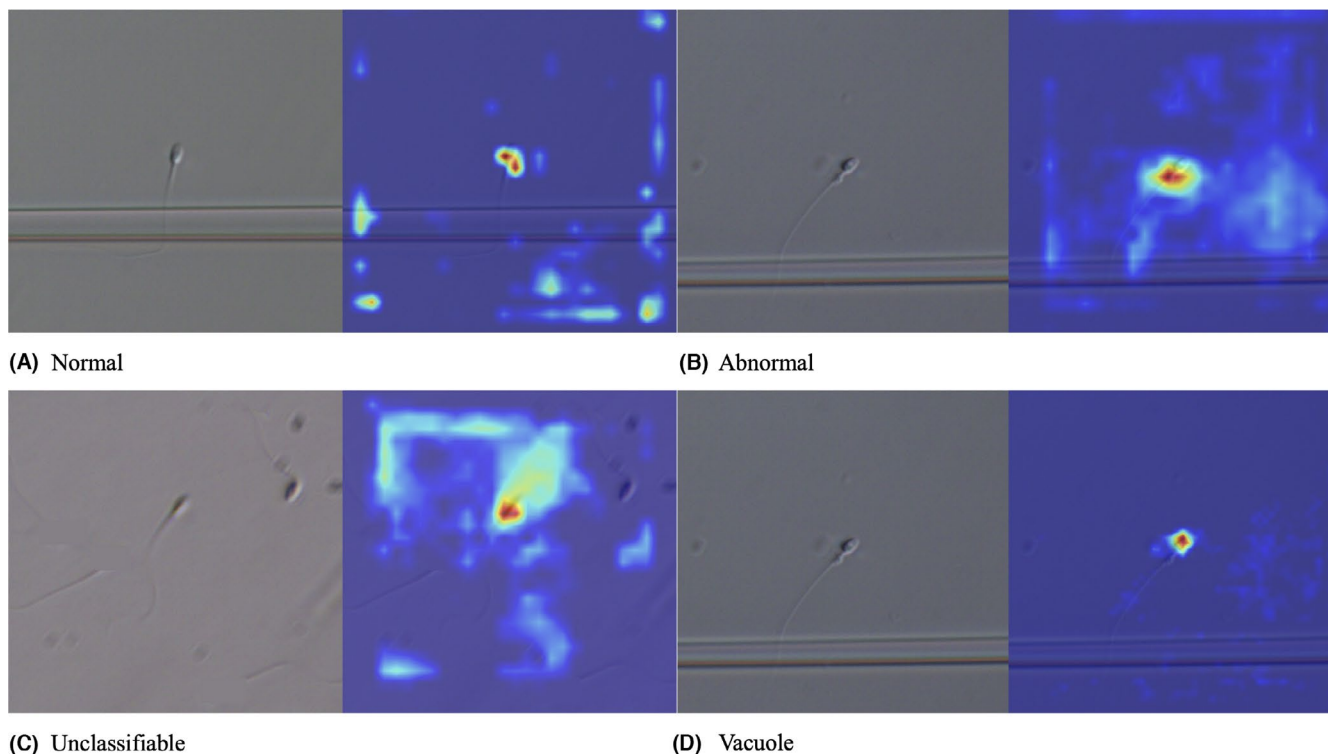**(A)** Normal **(B)** Abnormal **(C)** Unclassifiable **(D)** Vacuole

FIGURE 2 (color figure). Visual explanation of the proposed model generated by the Grad-CAM technique. When producing the inference results, the model pays attention to warm colors. Normal (A), abnormal (B), and unclassifiable (C) results are more commonly shown in the abstract layer L92, and the vacuole (D) shows the result at L104

small time-lag in normal moving image data, which is a little longer than the frame-switching time of 0.033 s.

3.2 | Results of the visual explanation

We attempted to make a visual explanation by showing a heat map created by the Grad-CAM technique adapted to this model. As shown in the figure, it was clarified that this model focuses on the part expected to make the judgment (Figure 2A–D). In the image showing an abnormality in the neck, the abnormality was predicted by focusing on the vicinity of the neck (Figure 2B). Normal sperm respond more widely from head to neck than abnormal sperm do (Figure 2A). The unclassifiable image reacts more extensively (Figure 2C). In sperm with vacuoles, a reaction was observed near the head in L104, which is a layer with a finer spatial resolution (Figure 2D). These findings are consistent with the points to which embryologists should pay attention.

TABLE 4 Results of the tracking performance of the proposed model

	n (%)
Unique object	51 (100)
Mostly tracked	40 (78.4)
Partially tracked	11 (21.6)
Mostly lost	0 (0)
False positive per frame	1.22 (2.4)
False negative per frame	2.18 (4.3)
ID switch	21[*]
MOTA (%)	84.37
MOTP	0.173

*Times for all frames.
Abbreviations: MOTA—multiple objects tracking accuracy; MOTP—multiple objects tracking precision.

3.3 | Tracking performance

In Table 2, we present the video and tracking performance of the results of the tracking AI detection (Video S3 and Table 4).

Among the 51 objects, 40 (78.4%) were mostly tracked, 11 (21.6%) were partially tracked, and 0 (0%) were mostly lost. There were 21 ID switches, indicating that tracking was interrupted and recognized as another sperm. The numbers of false positives and false negatives for the entire video were 185 and 332, with 1.22 (2.4%) and 2.18 (4.3%) per frame, respectively. With this level of tracking performance, it can be used in clinical settings without any discomfort. To improve the accuracy of tracking, we annotated objects that were almost out of focus when creating the true correct answer, so there were many false negatives. Most of these were objects that this model was unable to recognize as sperm because they were too out of focus. MOTA and MOTP were 84.4% and 0.173, respectively. The high tracking performance means that it is possible to display sperm motility and the trajectory at the same time as the morphological evaluation.

4 | DISCUSSION

We have shown that the proposed model makes short-time predictions for normal and abnormal sperm with sensitivities of 0.794 and 0.881 and PPVs of 0.689 and 0.853, respectively. It also demonstrated high tracking performance for video data that are easily out of focus. This high performance made it possible to evaluate the morphology of sperm while tracking, as carried out by embryologists during ICSI, making it the model closest to the clinical application at this time. In addition, by performing morphological evaluation and tracking simultaneously, it became possible to acquire time-series data of one sperm. This makes it easy to acquire and annotate sperm image data and helps solve the proposition of improving the pregnancy rate by linking the evaluation of sperm morphology with clinical information.

The proposed AI model may assist in detecting abnormal sperms in a clinical setting, reducing the workload of embryologists, and complementing subjectivity.

4.1 | Model performance of morphological assessment

Because we used the JDS in this study, the original unstained sperm image data set; it makes sense to compare its performance with other studies using the MHSMA data set, which also collects unstained sperm. However, a simple comparison would not make sense because of the different natures of these data sets.

The MHSMA data set, which consists of pictures of approximately 1,540 unstained sperm taken at low magnification of $400\times$ or $600\times$, is provided with expert judgment as to whether the head, acrosome, vacuoles, neck, and tail are normal or abnormal.[11] These

pictures are taken in a focused state and in a direction in which it can be determined whether the picture is normal or abnormal (mostly in a direction in which the maximum width can be measured). Although the data are slight, it also includes pictures that have been judged despite inappropriate orientation or angle for the classification.

The state-of-art model for the MHSMA data set at this time was reported by Abbasi et al,[11] who used a method called the "deep transfer learning technique" to address the lack of training data in MHSMA.[19] In this technique, the selected parameters of a neural network trained earlier on one or more data sets to solve source problems will be retrained on a new data set to tackle the target problem.[11] In the model proposed by Abbasi et al, the normal PPVs were very high at 93.4%, 91.8%, and 97.7% for each of the sperm acrosome, head, and vacuole, respectively, whereas the abnormal PPVs were 43.7%, 63.0%, and 65.8%, which were lower than the model used in this study. Those authors excluded the sperm neck and tail from the evaluation because they are easy to evaluate by experts and do not require assistance from an AI model. In addition, as the authors pointed out, the low variation in abnormal sperm in the MHSMA data set is considered to be one of the factors that do not improve the PPV for abnormal sperm. It is considered that because our proposed model was trained with a data set with more variations for cervical abnormalities, it had a higher PPV for abnormal sperm. In addition, the MHSMA data set consists of sperm with only good orientations that can be easily predicted and well-focused, and thus, the model that trained such data may not be able to recognize it as a sperm in clinical practice, and the frequency of false positives and false negatives are expected to increase. In our study, objects that could not be determined were classified as unclassifiable, which makes it possible that abnormal and normal characteristics became apparent.

The AI model developed in this study evaluated images under an inverted microscope in a short time, and its sensitivity and PPV for abnormal sperms were very high. This may help detect abnormal sperms in clinical settings, potentially reducing the workload of embryologists and complementing subjectivity.

4.2 | Model performance of tracking

There are several reports on tracking motile sperm using deep learning,[20–23] of which the latest and highest performing ones were reported by Somasundaram et al.[12] The sperm with the fastest movement speed was detected in 1.12 se, and the error rate was 2.31 with high accuracy. Those authors dealt with vertical defocus by using a cover glass to drop unstained sperm. A cover glass cannot be used when performing ICSI in the clinical setting, because an inverted microscope is used. If the cover glass is not used, it is difficult to focus because of the movement of sperm in the Z direction (perpendicular to the surface of the stage of the microscope). In this study, by creating a classification called "unclassifiable," we found it was possible to deal with a lack of focus and perform tracking. As a result, 40 of 51 (78%) sperm in the evaluated video were mostly tracked, and MOTA and MOTP were 84.4% and 0.173, respectively, making it possible

to evaluate sperm motility with high tracking performance. In the tracking competition for multiple objects (https://motchallenge.net/results/MOT20) held in 2020, the top models were MOTA 58% and MOTP 0.20, which indicate that the performance of this model is high. High-precision tracking of sperms, which are easily out of focus under an inverted microscope, may be practically useful because it enables velocity calculation. The short evaluation time of 0.246 s is also crucial for evaluating sperms in motion in clinical situations.

4.3 | About the setting of ground truth

At present, the fact that the "ground truth" is the judgment by the embryologist remains a problem. Sperm morphology is continuously present in varying degrees, from ideal normal sperm morphology, as shown in the literature, to apparently abnormal sperm. Therefore, setting a clear threshold between normal and abnormal is difficult.[24] In clinical practice, judgments differ to some extent depending on the semen findings of individual embryologists and the status of their patients. This ambiguity has also become a problem when defining the ground truth, and in some cases, the ground truth was reevaluated by the prediction of the AI model. This model is a digitized version of the sperm assessment technique of one embryologist. Assuming that the judgment criteria of each embryologist are potentially different, this algorithm cannot be used as it is in other facilities. However, it is considered possible to create a new algorithm suitable for the facility by changing annotation for a new judgment criterion. The AI prediction is just a prediction. Although this prediction is useful for detecting morphological abnormalities, there is still a need for the embryologist to make the final decision whether or not to use sperm for ICSI.

4.4 | How to use the proposed model in clinical situations

An embryologist's judgment can now be stored using such image recognition AI. Displaying the judgment results on screen may be effective in situations where the embryologists will be educated. It is also useful when determining the sperm selection criteria for each facility. The ability to detect abnormalities with a high degree of accuracy may also reduce the workload of embryologists. Furthermore, it may be possible to create an AI model that predicts fertilization and pregnancy by linking this model with clinical information such as the presence or absence of fertilization and pregnancy in addition to features extracted by an AI model capable of evaluating oocyte morphology.

However, there remains a concern that the introduction of AI models into daily clinical practice may itself reduce work efficiency. This is because once AI is introduced into clinical practice, the task of confirming the AI's judgment will be added to the usual workflow of performing ICSI. Embryologists will need to look away from the inverted microscope once to check the AI decision on a separate monitor. To improve the work efficiency loss due to the one-step addition, it is necessary to place a monitor displaying the AI results near the inverted microscope and increase familiarity of embryologists with the workflow using AI.

5 | CONCLUSIONS

In this study, we created the JSD, which is a high-quality data set acquired in accordance with the work process performed by embryologists during ICSI in clinical practice. Moreover, based on this data set, we created a new model that performs high-accuracy morphological prediction and tracking in short time using YOLOv3. This is the first report of a model that has high tracking performance and can perform the morphological assessment for video data that are difficult to focus. Rather than competing on the performance of AI models on a given data set, this study has been conducted in a practical clinical setting; to the best of our knowledge, this has not been reported before. Morphological evaluation while tracking and the short evaluation time may be useful in clinical settings to assist in the detection of abnormal sperms and reduce the burden on embryologists. In addition, by performing the morphological evaluation and tracking at the same time, it became possible to acquire time-series data of one sperm. This makes it easy to acquire and annotate sperm image data and helps to solve the proposition of improving the pregnancy rate by linking the evaluation of sperm morphology with clinical information. Future studies on highly practical models are expected.

CONFLICT OF INTEREST

Aikou Okamoto funded the collaborative research expenses through Olympus Corporation. Toshiyuki Hattori is paid by Olympus. Shinji Nakazawa is paid by LPIXEL Inc. Olympus Corporation paid R&D commission to LPIXEL Inc. The other authors declare that they have no known competing financial interests or personal relationships that influenced the work reported in this paper.

ETHICAL APPROVAL

This study was approved by the Institutional Review Board of Jikei University School of Medicine (IRB No. 30–086 9107).

HUMAN RIGHTS STATEMENTS AND INFORMED CONSENT

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation

(institutional and national) and with the Helsinki Declaration of 1964 and its later amendments. Informed consent was obtained from all patients for being included in the study.

ORCID

Takuma Sato https://orcid.org/0000-0002-9743-1004

Hiroshi Kishi https://orcid.org/0000-0003-1562-410X

REFERENCES

1. Isidori A, Latini M, Romanelli F. Treatment of male infertility. *Contraception*. 2005;72:314-318.
2. Levine H, Jørgensen N, Martino-Andrade A, et al. Temporal trends in sperm count: a systematic review and meta-regression analysis. *Hum Reprod Update*. 2017;23:646-659.
3. World Health Organization. *WHO Laboratory Manual For The Examination And Processing Of Human Semen*, 5th ed. World Health Organization; 2010.
4. Menkveld R, Stander FS, Kotze TJ, Kruger TF, van Zyl JA. The evaluation of morphological characteristics of human spermatozoa according to stricter criteria. *Hum Reprod*. 1990;5:586-592.
5. Hicks SA, Andersen JM, Witczak O, et al. Machine learning-based analysis of sperm videos and participant data for male fertility prediction. *Sci Rep*. 2019;9:16770.
6. Riordon J, McCallum C, Sinton D. Deep learning for the classification of human sperm. *Comput Biol Med*. 2019;111:103342.
7. Ilhan HO, Serbes G, Aydin N. Automated sperm morphology analysis approach using a directional masking technique. *Comput Biol Med*. 2020;122:103845.
8. Ilhan HO, Sigirci IO, Serbes G, Aydin N. A fully automated hybrid human sperm detection and classification system based on mobilenet and the performance comparison with conventional methods. *Med Biol Eng Comput*. 2020;58:1047-1068.
9. Iqbal I, Mustafa G, Ma J. Deep learning-based morphological classification of human sperm heads. *Diagnostics (Basel)*. 2020;10:325.
10. Javadi S, Mirroshandel SA. A novel deep learning method for automatic assessment of human sperm images. *Comput Biol Med*. 2019;109:182-194.
11. Abbasi A, Miahi E, Mirroshandel SA. Effect of deep transfer and multi-task learning on sperm abnormality detection. *Comput Biol Med*. 2021;128:104121.
12. Somasundaram D, Nirmala M. Faster region convolutional neural network and semen tracking algorithm for sperm analysis. *Comput Methods Programs Biomed*. 2021;200:105918.
13. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. *arXiv*. preprint 2016;arXiv:1506.02640.:
14. Redmon J, Farhadi A. YOLOv3: an incremental improvement. *arXiv*. preprint 2018; arXiv:1804.02767.
15. Lin T, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *arXiv*. preprint 2018;arXiv:1708.02002.
16. Ramprasaath R, Selvaraju MC, Das A, et al. Visual explanations from deep networks via gradient-based localization. *arXiv*. preprint 2019;arXiv:1610.02391.
17. Bewley A, Zongyuan G, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. *arXiv*. preprint 2017;arXiv:1602.00763v2.
18. Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP J Image Video Process*. 2008;2008:1-10.
19. Tan C, Sun F, Kong T, Zang W, Yanf C, Liu C. A survey on deep transfer learning. *arXiv*. preprint 2018;arXir:1808.01974v1.:
20. Mostajer Kheirkhah F, Sadegh Mohammadi HR, Shahverdi A. Modified histogram-based segmentation and adaptive distance tracking of sperm cells image sequences. *Comput Methods Programs Biomed*. 2018;154:173-182.
21. Li Q, Chen X, Zhang H, et al. Automatic human spermatozoa detection in microscopic video streams based on OpenCV. Biomed Eng. *Inform*. 2012;224-227.
22. Mohammadreza Ravanfar LA, Moradi MH, Fazel-Rezai R. Occlusion robust low-contrast sperm tracking using switchable weight particle filtering. *Adv Sex Med*. 2014;4:42-54.
23. Mostajer Kheirkhah F, Sadegh Mohammadi HR, Shahverdi A. Efficient and robust segmentation and tracking of sperm cells in microscopic image sequences. *IET Comput vis*. 2019;13:489-499.
24. Susan A, Rothmann AMB. *Sperm morphology*. 2nd ed. Encyclopedia of Reproduction; 2018: 85-95.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.