Received: 28 January 2025 | Accepted: 3 April 2025

DOI: 10.1002/rmb2.12649

Reproductive Medicine and Biology

WILEY

ORIGINAL ARTICLE

A machine learning model for predicting fertilization following short-term insemination using embryo images

Masato Saito[1,2] | Hirofumi Haraguchi[1] | Ikumi Nakajima[1] | Shinya Fukuda[1] | Chenghua Zhu[1] | Norio Masuya[1] | Kazunori Matsumoto[1] | Yuya Yoshikawa[3] | Tomoki Tanaka[1] | Satoshi Kishigami[2,4] | Leona Matsumoto[1]

[1]Matsumoto Ladies IVF Clinic, Tokyo, Japan

[2]Department of Integrated Applied Life Science, University of Yamanashi, Yamanashi, Japan

[3]Software Technology and Artificial Intelligence Research Laboratory, Chiba Institute of Technology, Chiba, Japan

[4]Center for advanced Assisted Reproductive Technologies, University of Yamanashi, Yamanashi, Japan

Correspondence
Hirofumi Haraguchi, Matsumoto Ladies IVF Clinic, Rokumarugate Ikebukuro Bld., 1-13-6 Higashiikebukuro, Toshimaku, Tokyo 170-0013, Japan.
Email: haraguchi-tky@umin.ac.jp

Abstract

Purpose: This study established a machine learning model (MLM) trained on embryo images to predict fertilization following short-term insemination for early rescue ICSI and compared its predictive performance with the embryologist's manual classification.

Methods: Embryo images at 4.5 and 8 h post-insemination were preprocessed into vectors using ResNet50. The Light Gradient Boosting Machine (Light GBM) was employed for training vectors. Fertilization in the test dataset was assessed by MLM, with seven senior and 11 junior embryologists. Predictive metrics were analyzed using repeated measures ANOVA and paired *t*-tests.

Results: Comparing MLM, senior embryologists, and junior embryologists, significant differences were observed in accuracy (0.71 ± 0.01, 0.75 ± 0.05, 0.61 ± 0.05), recall (0.84 ± 0.02, 0.84 ± 0.10, 0.61 ± 0.07), F1-score (0.78 ± 0.01, 0.81 ± 0.04, 0.66 ± 0.04), and area under the curve ($0.73 \pm 0.0\ 3$, 0.73 ± 0.06, 0.61 ± 0.07), the MLM outperforming junior embryologists with <1 year of experience. No significant differences were observed between the MLM and senior embryologists with over 5 years of experience.

Conclusions: MLM can effectively predict fertilization following short-term insemination by analyzing cytoplasmic changes in images. These results underscore the potential to enhance clinical decision-making and improve patient outcomes.

KEYWORDS

cytoplasm, early rescue ICSI, fertilization, machine learning, short-term insemination

1 | INTRODUCTION

In vitro fertilization (IVF) treatment is a form of assisted reproductive technology, with conventional IVF (c-IVF) typically used as the initial method of fertilization.[1] However, approximately 5%–20% of patients experience complete fertilization failure (CFF) post c-IVF.[2–4] CFF results in the cancellation of the treatment cycle, thereby requiring intracytoplasmic sperm injection (ICSI) for the subsequent cycle. CFF imposes considerable physical and financial burdens on patients. To circumvent CFF, rescue ICSI (r-ICSI), which

involves performing ICSI 21–33 h after oocyte retrieval, has been reported.[5] The prolonged interval between oocyte retrieval and fertilization compromises the oocyte quality, leading to the failure of r-ICSI. On the other hand, early r-ICSI, such as within 8 h, helps minimize oocyte damage and decline in quality by performing ICSI on unfertilized oocytes following short-term insemination. Previous studies have demonstrated that early r-ICSI can potentially prevent oocyte degradation and achieve higher pregnancy rates.[6] Therefore, the decision to perform early r-ICSI should be made by embryologists 4–8 h after insemination.[6–9] Early r-ICSI is performed on oocytes that do not exhibit fertilization indicators, such as pronuclei, extrusion of the second polar body, cytoplasmic wave, or fertilization cone.[10] The second polar body, which is a by-product of meiosis, is extruded within 4 h after fertilization. Additionally, pronuclei appear as early as 8 h after insemination. These can be readily used by embryologists to classify fertilization.[8,11] Conversely, if polar bodies are fragmented and pronuclei have not appeared, the embryologist's classifications based solely on the cytoplasmic wave and fertilization cone are challenging due to their subjective nature and lack of objectivity. Furthermore, the precision of fertilization classification is contingent upon the expertise of embryologists. Unnecessary ICSI on fertilized embryos should be avoided to prevent the risk of polyspermy. Errors in fertilization classification not only risk the oversight of unfertilized oocytes but also elevate the potential for polyspermy.[12]

According to a previous study, the spindle fiber observation using an inverted microscope made it possible to classify fertilization in oocytes with fragmented polar bodies. The technique allows classification without depending on the second polar body. On the other hand, the spindle fibers observation has the problem of causing unavoidable environmental damage to the embryo when performed outside the incubator.[11] Furthermore, classification by observing spindle fibers requires trained embryologists and specialized equipment for differential interference contrast microscopy. Thus, spindle fiber observation requires the purchase of expensive equipment and increases the routine workload for the observation.

Machine learning (ML) has been shown to be an alternative observational technique as it excels at identifying intricate patterns and hidden correlations through data-driven learning. ML generates data-driven predictions without relying on experience or intuition.[13] The application of ML to fertilization classification is expected to provide objective and reproducible outcomes, independent of the embryologist's experience. The feature patterns learned by ML offer a unique perspective that differs from traditional human classification and provides valuable alternative insights. However, studies that implement ML for fertilization classification and evaluation of predictive performance have not yet been reported.

This study aims to establish an ML model (MLM) for the fertilization classification of embryos following short-term insemination, trained on embryo images, and to compare the predictive performance of the MLM and the embryologist's manual classification.

2 | MATERIALS AND METHODS

2.1 | Study criteria

This retrospective study was approved by the ethical committee of the Japanese Institution for Standardizing Assisted Reproductive Technology (JISART; 2024-21). This study used data collected from January to October 2021. Data were gathered from short-term insemination cycles following oocyte retrieval. Fertilization was assessed by two embryologists observing the pronuclei from their appearance to their disappearance to avoid missing early disappearance or delayed appearance of pronuclei. The embryos were classified into two pronuclei (2PN) and no pronuclei (0PN) groups, while one pronuclei (1PN) and three pronuclei (3PN) were excluded from this study due to their limited numbers. This study was non-interventional, and the results were not employed in any manner to influence treatment decisions.

2.2 | Ovarian stimulation, oocyte retrieval, and IVF procedure

Patients underwent ovarian stimulation using the Progestin-Primed Ovarian Stimulation (PPOS) protocol. The optimal dosage of follicle-stimulating hormone (FSH) was determined based on the serum concentrations of basal FSH and anti-Müllerian hormone (AMH), in conjunction with the patient's age. FSH dosage ranged from 150 to 300 units per day. Additionally, medroxyprogesterone acetate (MPA) at a dosage of 5–10 mg or dydrogesterone at a dosage of 20 mg/day, was administered. Oocyte maturation was induced using a GnRH agonist when the leading follicle diameter, measured via transvaginal ultrasound, was ≥18 mm. The procedure involved oocyte retrieval 34–35 h post-induction, performed under transvaginal ultrasound guidance. Semen samples were processed using the density gradient centrifugation method with Extra Sperm Selection™ (ORIZURU ART Family, Kyoto, Japan) by centrifugation at 400 *g* for 20 min. The recovered pellet was suspended in 4 mL of Gx-IVF™ (Vitrolife AB, Gothenburg, Sweden) and centrifuged for 5 min at 300 *g*. Finally, the pellet was resuspended again in 0.5 mL of Gx-IVF™ (Vitrolife AB). 1×10^5 motile sperm were used for insemination in 1 mL of Gx-IVF™ (Vitrolife AB) containing cumulus-oocyte complexes. The sperm were co-cultured with the oocytes for 4.5 h. Following co-culture, cumulus cells underwent denudation, and the resulting embryos were cultured in an EmbryoScope™ time-lapse incubator (Vitrolife AB).

2.3 | Embryo imaging and preprocessing

Embryo images were captured at 4.5 and 8 h post-insemination using a time-lapse incubator. The images were resized from 800 × 800 to 224 × 224 pixels. The circular Hough transform algorithm was used to detect the cytoplasm, thereby minimizing the

impact of noise, and the area outside the circle was masked in black. The RGB values of each pixel in the embryo images were centralized by subtracting the mean RGB value of the entire image. Subsequently, the standard deviation of the RGB values of the entire image was calculated, and each pixel value was normalized by dividing it by the standard deviation.

2.4 | MLM

Initially, 10% of the images of 878 embryos were randomly selected as the test dataset, while the remaining 90% were designated as the training dataset. The weights of the pretrained convolutional neural network, ResNet50, were fixed and utilized as feature extractors. The preprocessed embryo images at 4.5 and 8 h were input into ResNet50, converting them into 2048-dimensional vectors. The vectors from the 4.5 and 8 h images were concatenated. Consequently, the two images were transformed into 4096-dimensional vectors (Figure 1). The training dataset, which was transformed into 4096-dimensional vectors, was partitioned by retaining 20% of the validation data. The Light Gradient Boosting Machine (Light GBM) analysis algorithm was employed, which is known for its rapid processing and efficient enhancement of its predictive function.[14] Compared to the standard gradient boosting tree algorithm, Light GBM enhances efficiency and accelerates computation by optimizing histograms, which improves segmenting features and conserves computational memory.[15] Hyperparameter tuning was performed using validation data by employing Bayesian optimization through the Optuna framework. The training dataset was divided into training and validation sets in a 4:1 ratio. The Optuna framework explored the following hyperparameters (range) involved in the training of Light GBM: the metric (binary_log-loss), lambda_l1 ($1e^{-8}$–10.0), lambda_l2 ($1e^{-8}$–10.0), num_leaves (2–256), feature_fraction (0.4–1.0), bagging_fraction (0.4–1.0), bagging_freq (1–7), and min_child_samples (5–100). Using the hyperparameters tuned by the Optuna framework, we evaluated the predictive performance with 5-fold cross-validation.

2.5 | Fertilization predictive performance

The trained Light GBM model was applied to predict the outcomes of the test dataset. Furthermore, 10 different random seeds were configured, and 10 predictions on the test dataset were executed. To determine the reliability of ML predictions, predictions made by embryologists were compared with the ML predictions. Eighteen embryologists had predicted whether the sperm had fertilized the oocyte in the test dataset. Embryologists with over 5 years of experience were classified as senior embryologists, and those with <1 year of experience as junior embryologists. Embryologists reviewed the movie of each embryo taken 4.5 to 8 h following short-term insemination and classified them as either 2PN or 0PN based on the comprehensive assessment of the appearance of the second polar body, cytoplasmic wave, and fertilization cone. The embryologist's review was performed in a blinded manner without knowledge of the true ratio in the test data set. The metrics evaluated for each prediction included accuracy, recall, specificity, precision, F1 score, area under the ROC curve (AUC), true positive ratio, false positive ratio, true negative ratio, and false negative ratio. Accuracy is the ratio of correctly predicted instances to the total instances, reflecting the

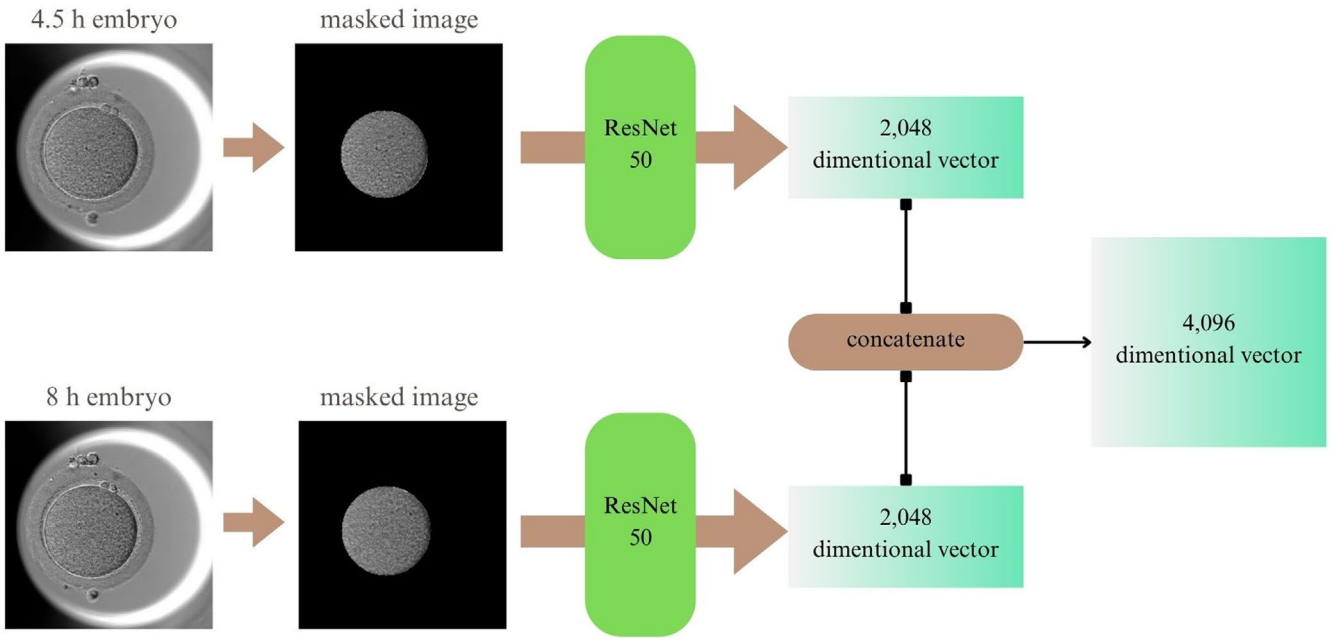

FIGURE 1 Flowchart depicting the preprocessing of embryo images captured at 4.5 and 8 h post-insemination, respectively. Hough circle transformation was used to mask the cytoplasmic area outside to minimize the effect of noise. The masked images were converted to 2048-dimensional vectors for input into ResNet50. Converted vectors were concatenated.

overall correctness of the model. Recall is the proportion of actual positive cases that were correctly identified by the model, indicating its ability to detect positive instances. Specificity is the proportion of actual negative cases that were correctly identified by the model, indicating its ability to detect negative instances. Precision is the proportion of predicted positive cases that were correctly identified, reflecting the accuracy of positive predictions. The F1 score is the harmonic mean of precision and recall, providing a single measure of the model's balance between precision and recall. The AUC is a performance measurement for classification models at various threshold settings, representing the degree of separability between classes. The confusion matrix metrics were calculated as a ratio of the number of instances in each category by the total number of instances. Furthermore, the error rate for each embryo in the test dataset was calculated to compare the error trends. A comparison was performed on embryos with an error rate of over 50% in either MLM or senior embryologist classification. The error rate was calculated as the ratio of misclassified cases to the total number of classifications. The LIME method visualizes how image regions contribute to the MLM prediction.[16] We used the LIME method to assess the validity of the classification by visualizing the critical cytoplasmic regions in an error rate of 0% embryos. Furthermore, to confirm the ability to adjust sensitivity, the predictive performance at different cutoff values of the MLM prediction score was evaluated.

2.6 | Statistical analyses

Repeated measures ANOVA was conducted on the predictions made by the MLM, junior, and senior embryologists. Paired *t*-tests were conducted when significant differences were found in the repeated measures ANOVA. Multiple comparisons were performed using the Bonferroni correction. A comparison of error rates was conducted using residual analysis. The significance level was set at $\alpha < 0.05$. Statistical analyses were performed with EZR (Saitama Medical Center, Jichi Medical University, Saitama, Japan).[17]

3 | RESULTS

In 230 cycles, 1491 oocytes underwent short-term insemination. Of these, 91 and 41 embryos were classified as 3PN and 1PN, while 391 and 90 immature and degenerated oocytes were excluded from the analysis, respectively. Ultimately, 547 and 331 embryos were identified as 2PN and 0PN, respectively, resulting in 878 embryos. The 0PN group included 23 arrested embryos at the second polar body stage. The 878 embryos were divided using a random test split method, with 790 embryos allocated to the training dataset and 88 embryos to the test dataset (Figure 2).

The Light GBM hyperparameters were tuned using the Optuna framework and adjusted based on 5-fold cross-validation. The tuning results were as follows: the metric (binary_error), lambda_l1 (9.209990901687287e-08), lambda_l2 (5.176175467675683e-08), num_leaves (5), feature_fraction (0.5284113215213745), bagging_fraction (0.6207237760711464), bagging_freq (7), and max_depth (4). The metric was changed from binary_log-loss to binary_error since binary_error had higher predictive performance. The performance consistency of the tuned MLM was confirmed using 5-fold cross-validation on the training dataset (Table 1).

Table 2 shows the predictive performance of the test dataset, as estimated by the MLM after 10 iterations, in comparison to the assessments of the seven senior and 11 junior embryologists, respectively. Repeated measures ANOVA revealed statistically significant differences in the metrics of accuracy, recall, F1-score, AUC, true positive, and false negative. In contrast, no significant differences were observed for the other metrics. Upon conducting multiple comparisons of the metrics for accuracy, recall, F1-score, AUC, true positive, and false negative using paired *t*-tests, junior embryologists exhibited significantly lower performance than other groups. No significant differences were observed between the MLM and the senior embryologists. The individual predictive performance is detailed in Table S1.

The comparison of 31 embryos with an error rate of over 50% in either the MLM or the senior embryologist classification showed a significant difference in 26 embryos within the test dataset using residual analysis. The MLM had a significantly lower error rate than the mean error rate of three groups for six embryos and a significantly higher error rate for 17 embryos. The error rate of the senior embryologist was significantly lower for nine embryos and higher for five embryos. The error rate of the junior embryologist was significantly lower for 11 embryos and higher for four embryos (Table 3).

The LIME method visualizes the contributions of image regions to the prediction score. Here, the green image regions indicate contributions to 2PN, while the red ones indicate contributions to 0PN. The presence of green regions within the cytoplasm indicates that MLM utilized cytoplasmic changes (Figure 3).

It was indicated that the predictive performance could be adjusted by adopting different cutoff values for the MLM prediction score on the training dataset. The true positive ratio and false positive ratio both showed a declining trend with increasing cutoff values. The true negative ratio and false negative ratio showed an increasing trend (Table S2).

4 | DISCUSSION

This study indicates that the MLM can predict fertilization following short-term insemination by analyzing cytoplasmic changes from embryo images. The predictive performance of the MLM was significantly higher than that of the junior embryologists, with no statistical difference observed compared to the senior embryologists. These findings underscore the potential of ML to enhance the reliability of fertilization prediction, which could lead to more informed clinical decisions.

Previous studies have completely relied on manual classification for the implementation of early r-ICSI.[18] Early r-ICSI permits clinical

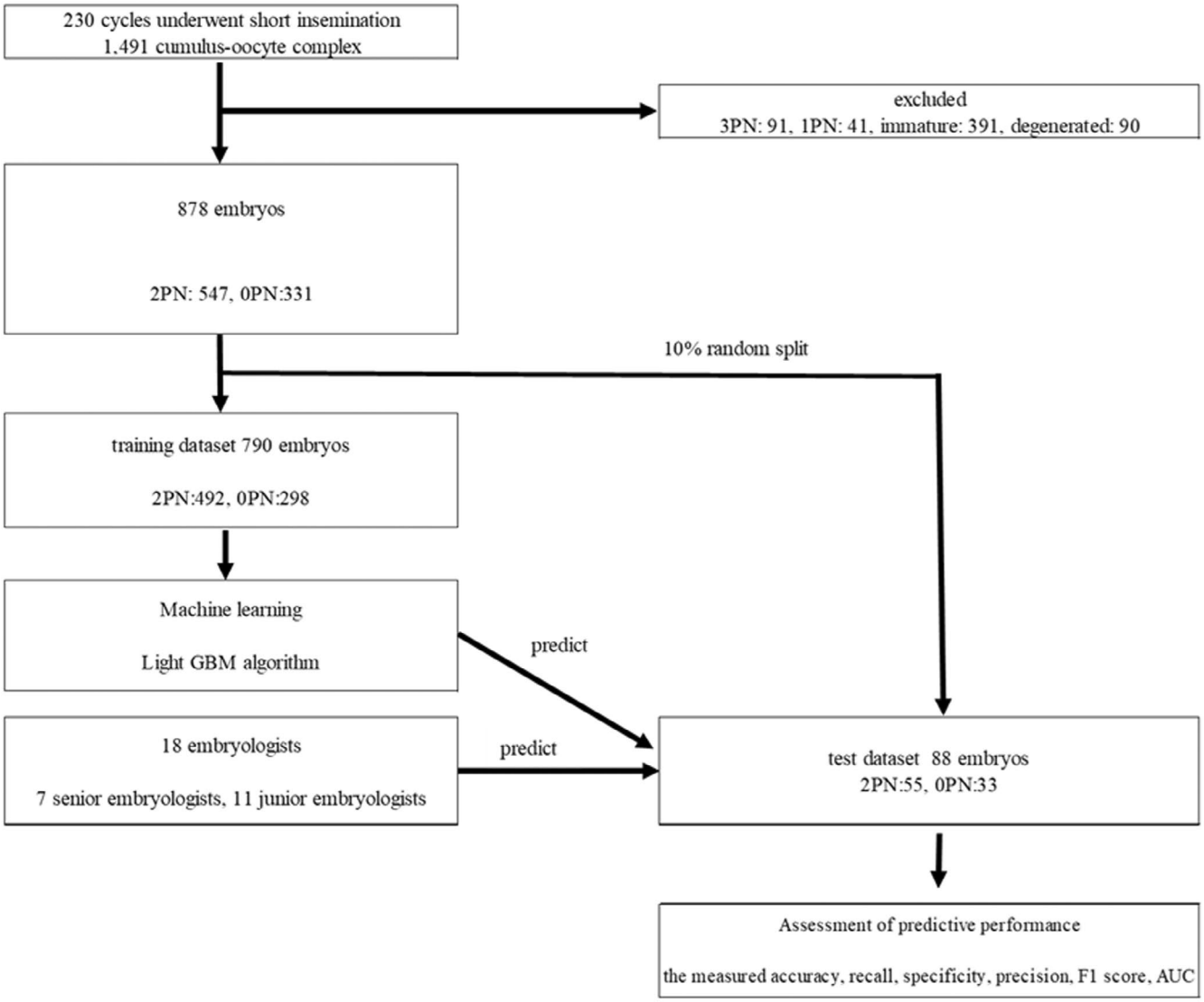

FIGURE 2 Flowchart depicting the study design and results of the subsequent analyses. The 878 embryos underwent short-term insemination were divided using a random test split method, with 790 embryos allocated to the training dataset and 88 embryos to the test dataset. The predictive performance of the machine learning model and 18 embryologists was evaluated using the 88 embryos in the test dataset. AUC, area under the ROC curve; Light GBM, Light Gradient Boosting Machine.

utilization of unfertilized oocytes following short-term insemination. The clinical utilization of unfertilized oocytes contributes to the prevention of CFF and a reduction in the number of oocyte retrievals. Reducing time to pregnancy significantly alleviates the psychological and economic burdens on patients.[19] Furthermore, previous research has indicated that early r-ICSI does not significantly differ from conventional ICSI in terms of clinical pregnancy rates, miscarriage rates, and neonatal outcomes.[20] The rates of congenital birth defects were also similar across the groups, suggesting the safety of this approach.[20,21] However, it is inherently associated with the increased risk of polyspermy owing to human error in fertilization classification. Embryologists' fertilization classification is highly dependent on their expertise and experience, leading to inevitable interindividual variability and susceptibility to human error.[11] 6–8 h post-insemination, extrusion of the second polar body is considered the most accurate fertilization indicator.[12,18] When the polar body assumes fragmentation or sperm entry is delayed beyond the typical 1–4 h, the accurate assessment of fertilization becomes challenging.[22] In such cases, it is necessary to rely on the presence of a fertilization cone or cytoplasmic wave. Nevertheless, the incidence of fertilization cones has been reported to be as low as 3.6%.[23] Since cytoplasmic wave typically occurs 2–3 h after extrusion of the second polar body, sufficient information is not obtained from the observation at the 8 h post-insemination point.[23] The fertilization classification based on limited information contributes to oversight of unfertilized oocytes and the occurrence of iatrogenic polyspermy. Conversely, this study demonstrated that the ML-based approach provides a more consistent and reproducible evaluation. The MLM can efficiently perform classification by learning complex correlations from images.[24] The consistency confirmed through 5-fold

TABLE 1 Predictive performance of the machine learning model using 5-fold cross-validation on the training dataset.

Fold	1	2	3	4	5	Mean ± SD
Accuracy	0.66	0.74	0.72	0.72	0.70	0.71 ± 0.03
Recall	0.77	0.88	0.89	0.94	0.78	0.85 ± 0.07
Specificity	0.47	0.51	0.43	0.37	0.57	0.47 ± 0.07
Precision	0.71	0.75	0.72	0.71	0.75	0.73 ± 0.02
F1 score	0.74	0.81	0.79	0.81	0.76	0.78 ± 0.03
AUC	0.74	0.78	0.74	0.77	0.74	0.75 ± 0.02
True positive ratio	0.48	0.55	0.55	0.58	0.48	0.53 ± 0.04
False positive ratio	0.20	0.18	0.22	0.24	0.16	0.20 ± 0.03
True negative ratio	0.18	0.19	0.16	0.14	0.22	0.18 ± 0.03
False negative ratio	0.15	0.08	0.07	0.04	0.14	0.09 ± 0.04

Note: The predictive performance of the machine learning model (MLM) was evaluated using 5-fold cross-validation on the training dataset ($n = 790$). The cutoff value for the prediction score using MLM was set at 0.5. The confusion matrix metrics were calculated as a ratio by the number of instances in each category by the total number of instances.

Abbreviation: AUC, area under the ROC curve.

TABLE 2 Comparative predictive performance of machine learning model and embryologist on test dataset.

Metric	Machine learning model (mean ± SD)	Senior embryologist (mean ± SD)	Junior embryologist (mean ± SD)	rANOVA *p*-value
Accuracy	0.71 ± 0.01	0.75 ± 0.05	0.61 ± 0.05*	<0.01
Recall	0.84 ± 0.02	0.84 ± 0.10	0.61 ± 0.07*	<0.01
Specificity	0.50 ± 0.03	0.61 ± 0.15	0.62 ± 0.17	0.14
Precision	0.74 ± 0.01	0.79 ± 0.06	0.73 ± 0.07	0.22
F1 score	0.78 ± 0.01	0.81 ± 0.04	0.66 ± 0.04*	<0.01
AUC	0.73 ± 0.03	0.73 ± 0.06	0.61 ± 0.07*	<0.01
True positive ratio	0.53 ± 0.01	0.52 ± 0.06	0.38 ± 0.04*	<0.01
False positive ratio	0.19 ± 0.01	0.14 ± 0.05	0.14 ± 0.06	0.16
True negative ratio	0.19 ± 0.01	0.23 ± 0.05	0.23 ± 0.06	0.14
False negative ratio	0.10 ± 0.01	0.10 ± 0.06	0.25 ± 0.04*	<0.01

Note: Repeated measures ANOVA (rANOVA) was used to compare the mean of each metric for the machine learning model, junior embryologists, and senior embryologists. Paired *t*-tests were used for multiple comparisons of metrics showing significant differences in the rANOVA. Superscripts (*) indicate junior embryologists have significant differences from both the machine learning model (MLM) and senior embryologists ($p < 0.01$). There was no statistically significant difference between the machine learning model and senior embryologist in each metric. The cutoff value for the prediction score using MLM was set at 0.5. The confusion matrix metrics were calculated as a ratio by the number of instances in each category by the total number of instances.

Abbreviation: AUC, area under the ROC curve.

cross-validation demonstrates the robustness and generalizability of the data.[25] Maintaining consistent predictive performance with different datasets underscores the advanced learning capabilities of the ML algorithm and highlights its potential as a valuable support tool in clinical practice. The MLM could serve as an effective adjunct to human expertise and enhance the reliability and consistency of fertilization classification. Fertilization classification applying ML contributes to the conservation of human resources and the reduction of errors in clinical practice.

In this study, the MLM was trained to learn and predict temporal changes in the cytoplasm from embryo images. The LIME method was used to visualize how image regions contribute to the MLM prediction, but it was difficult to interpret why the highlighted regions were relevant to the fertilization classification with the naked eye. Accompanying fertilization, cytoplasmic changes include variations in lipid droplets (LDs) and the cytoskeleton.[26,27] Lipids are fundamental cellular components involved in cell construction, metabolism, and regulation, with LDs serving as pivotal intracellular structures for energy storage and membrane synthesis.[28] It has been reported that LD size increases during the transition from maturation to development.[26] In addition, the cytoskeleton undergoes dramatic changes during fertilization. The actin cytoskeleton, particularly crucial for sperm entry and subsequent calcium signal transduction, undergoes significant reorganization.[27] Microtubules play an essential role in organizing mitotic spindles and promoting pronuclear formation, with the sperm centrosome contributing to the microtubule dynamics.[29] The changes in LDs are in the order of micrometers, and cytoskeletal alterations are extremely challenging

TABLE 3 Comparison of error rate between senior embryologist classification and machine learning model classification in embryos with over 50% error rate by either machine learning model or senior embryologist classification.

Embryo no.	Fertilization	MLM error rate ($n = 10$)	Senior embryologist error rate ($n = 7$)	Junior embryologist error rate ($n = 11$)
43	0PN	0.0%*	57.1%*	27.3%
82	0PN	0.0%*	57.1%	36.4%
55	2PN	0.0%*	57.1%	72.7%*
72	2PN	0.0%*	57.1%	81.8%*
42	2PN	0.0%*	85.7%*	81.8%*
31	0PN	0.0%*	100.0%*	81.8%*
40	0PN	10.0%	57.1%*	9.1%
1	2PN	60.0%	14.3%*	63.6%
84	0PN	60.0%	28.6%	18.2%
74	0PN	60.0%	85.7%*	27.3%*
12	0PN	60.0%	85.7%	81.8%
36	2PN	70.0%	14.3%	45.5%
76	2PN	80.0%*	0.0%*	9.1%*
61	0PN	80.0%*	42.9%	0.0%*
4	0PN	80.0%*	42.9%	18.2%*
9	0PN	80.0%	57.1%	54.5%
71	0PN	90.0%*	14.3%	18.2%*
59	2PN	90.0%*	71.4%	18.2%*
53	0PN	100.0%*	0.0%*	18.2%*
73	2PN	100.0%*	0.0%*	18.2%*
77	2PN	100.0%*	0.0%*	45.5%
54	0PN	100.0%*	14.3%*	45.5%
16	2PN	100.0%*	14.3%*	54.5%
49	0PN	100.0%*	14.3%*	63.6%
68	0PN	100.0%*	28.6%	36.4%
33	0PN	100.0%*	28.6%*	54.5%
25	0PN	100.0%*	42.9%	27.3%*
27	0PN	100.0%*	42.9%	54.5%
66	0PN	100.0%*	57.1%	18.2%*
50	0PN	100.0%	57.1%	63.6%
48	0PN	100.0%*	85.7%	45.5%*

Note: A comparison was performed on embryos with an error rate of over 50% in the machine learning model (MLM) and senior embryologist classification. Sorted by MLM error rate in ascending order. The error rate was calculated as the ratio of misclassified cases to the total number of classifications. Residual analysis was performed on the error rates in three groups.

*Significant difference between the error rate and the mean error rate of the three groups ($p < 0.05$).

to discern with the naked eye. The ML algorithm in this study is suggested to learn the differences imperceptible to human observation and leverage them for predictions.

Within this investigation, the MLM outperformed junior embryologists significantly in terms of accuracy, recall, F1-score, and AUC but showed no statistical difference when compared with senior embryologists. These findings indicate that the MLM exhibits a superior predictive performance compared to that of junior embryologists. The MLM suggests that it could be utilized as an effective educational tool, providing junior embryologists with additional resources for training and skill development.[30] Regarding recall, ML's ability to detect critical cases with minimal omission is evident, which is particularly advantageous in clinical settings where the identification of crucial cases is imperative.[31] The significant difference between the MLM and junior embryologists indicates that the overall performance of the MLM surpasses that of junior embryologists, and the MLM's detection capabilities can compensate for the embryologist's experience and contribute to improved diagnostic reliability, demonstrating the broader applicability of ML algorithms in fertilization classification by embryologists. The F1-score demonstrates that the

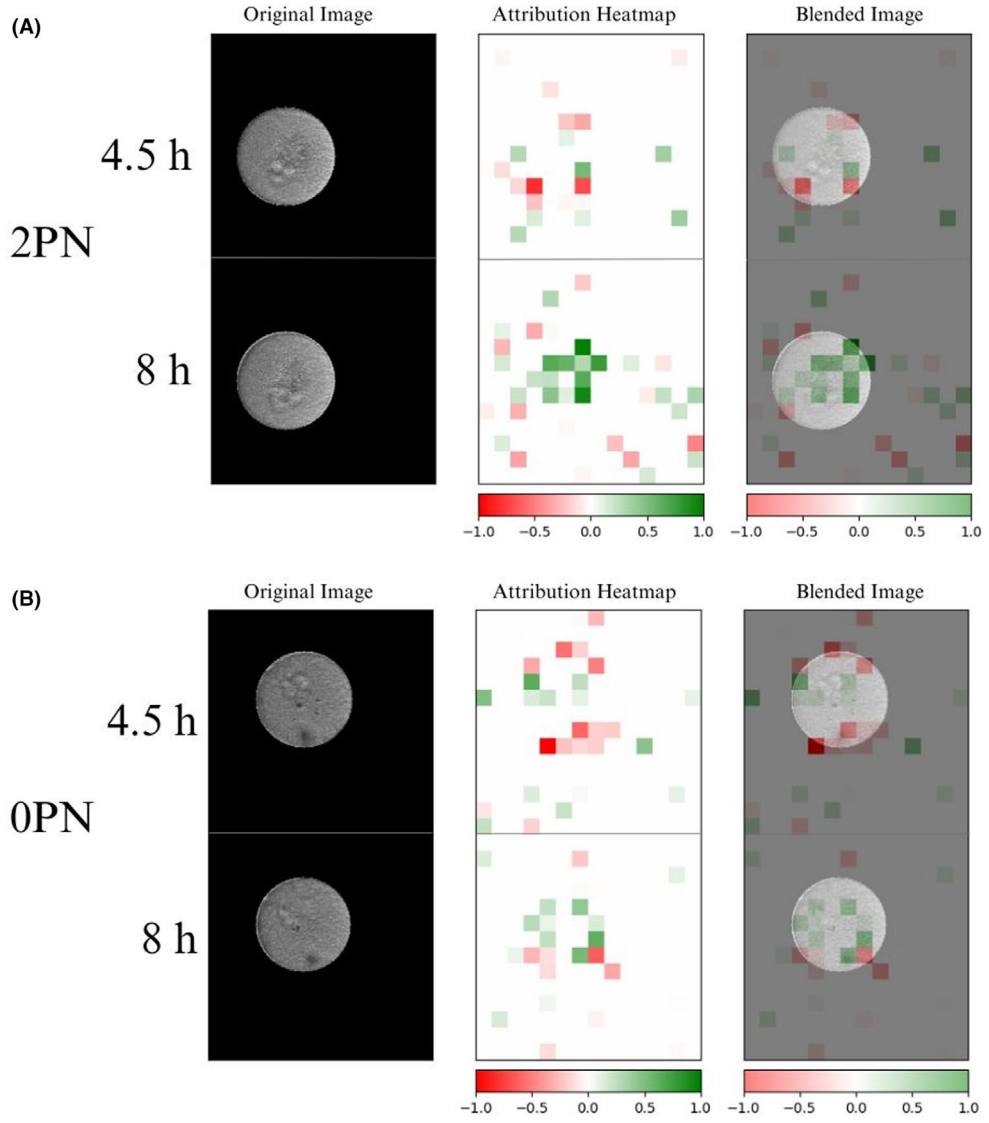

FIGURE 3 The contributions of image regions to the prediction score in the two input images were calculated using the LIME method. The LIME method indicates that the green regions contribute to 2PN and the red ones contribute to 0PN. The presence of regions with high contributions within the cytoplasm indicates that the MLM utilized cytoplasmic changes. (A) An example of a 2PN embryo from the test dataset with an error rate of 0% (Embryo No. 55). (B) An example of an 0PN embryo from the test dataset with an error rate of 0% (Embryo No. 82).

MLM maintains a balanced detection ability, highlighting both high precision and recall, which is paramount for clinical applications.[32] Moreover, the AUC reflects a well-balanced trade-off between true positive and false positive rates, providing a comprehensive measure of a model's predictive performance.[33] The significant difference in the true positive ratio and false negative ratio was potentially attributed to the experience level of the junior embryologists. The inability to detect cytoplasmic changes, due to their limited experiential knowledge, is considered to be the cause of misclassifying fertilized embryos as unfertilized. The MLM's that don't rely on experiential knowledge serve as valuable support tool for junior embryologists. The false negative ratio of fertilized embryos misclassified as unfertilized represents the risk of iatrogenic polyspermy in early r-ICSI. In addition to showing no significant difference in the false negative ratio compared to senior embryologists, the MLM can support the reduction of polyspermy by adjusting its sensitivity as needed. Although there was no significant difference, the specificity of junior and senior embryologists tended to be higher than that of the MLM. Additionally, the seven embryos with an error rate of over 50% by both MLM and senior embryologist classification were mostly 0PN and showed minimal cytoplasmic changes. The possibility that the MLM is less proficient in classifying unfertilized embryos compared to embryologists is due to over-detecting irrelevant cytoplasmic changes.

The spindle fiber observations that do not depend on polar bodies have been explored to reduce the incidence of polyspermy.[34] As spindle fiber observation requires the use of an inverted microscope, potential damage from environmental changes is unavoidable.[35]

The greater the number of embryos observed for the spindle, the greater the cumulative damage inflicted on the embryos. Conversely, the ML approach reduces potential damage by eliminating the need for embryo exposure to the external environment. Providing additional information to classify fertilization without transferring embryos from the incubator reduces the routine workload on embryologists. Seamless integration into the incubation process optimizes embryo quality and overall outcomes.

The MLM can achieve the same level of predictive performance as senior embryologists in predicting fertilization, marking a significant advancement in reproductive medicine. Analysis of the error rates for individual embryos in the test dataset showed that the MLM provided accurate classifications in embryos where embryologists tended to make classification mistakes. The six embryos that the MLM statistically had a lower error rate exhibited minimal cytoplasmic changes and polar body alterations. This finding supports that the MLM utilizes subtle cytoplasmic changes in embryos that are difficult for embryologists to recognize. The embryos with a statistically higher than the mean error rate of the three groups in the MLM had a greater number of cumulus cells attached to the zona pellucida. Cumulus cells overlapping the cytoplasm contribute to misclassification by introducing noise. There is potential for improving predictive performance by increasing the training data for embryos with overlapping cumulus cells and by removing these cells. Most of the embryos with statistically lower than the mean error rate of the three groups in the junior embryologist were 0PN. The junior embryologist classified the embryos as 0PN due to being less sensitive to cytoplasmic changes, whereas the MLM and senior embryologist over-detected cytoplasmic changes and classified them as 2PN. On the other hand, although there was no significant difference between the MLM and senior embryologists, the overall predictive performance of senior embryologists was better. Since senior embryologists perform classification using time-lapse movies, and the MLM relies on only two images, there could be a discrepancy in the amount of information available for learning. Incorporating time-lapse videos into MLM training is expected to further enhance its predictive performance. The results of this study highlight the sophisticated learning capabilities of the ML algorithm, indicating its broad applicability across various medical fields. Applying the ML algorithm to classification could enhance diagnostic and prognostic reliability across specialties in clinical practice.

In conclusion, this research revealed that the MLM can effectively predict fertilization following short-term insemination by learning cytoplasmic changes in embryo images from a different perspective of embryologists. The predictive performance of MLM was significantly superior to that of junior embryologists, with no statistical difference when compared with senior embryologists. These findings underscore the potential of ML to significantly enhance fertilization prediction performance, leading to more informed clinical decisions and improved patient outcomes. Despite these promising results, this study has limitations, including the specific conditions under which the data were collected. Without the inclusion of 1PN and 3PN in the training, classification accuracy was affected due to the untrained abnormal cytoplasmic changes. Incorporating abnormal patterns into the training data is important to maintain the consistency of classification accuracy. Future research should aim to replicate these findings in diverse clinical settings and enhance the robustness of prediction by adding learning data to resist noise.

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to the members of the Matsumoto Ladies IVF Clinic Laboratory and Yamanashi University Laboratory for their invaluable assistance with data collection.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

Masato Saito https://orcid.org/0009-0006-2437-6813
Satoshi Kishigami https://orcid.org/0000-0001-9447-5100

REFERENCES

1. Isikoglu M, Avci A, Kendirci Ceviren A, Aydınuraz B, Ata B. Conventional IVF revisited: is ICSI better for non-male factor infertility? Randomized controlled double blind study. J Gynecol Obstet Hum Reprod. 2021;50:101990.
2. Bhattacharya S, Hamilton MP, Shaaban M, Khalaf Y, Seddler M, Ghobara T, et al. Conventional in-vitro fertilisation versus intracytoplasmic sperm injection for the treatment of non-male-factor infertility: a randomised controlled trial. Lancet. 2001;357:2075–9.
3. Liu DY, Baker HW. Defective sperm-zona pellucida interaction: a major cause of failure of fertilization in clinical in-vitro fertilization. Hum Reprod. 2000;15(3):702–8. https://doi.org/10.1093/humrep/15.3.702
4. Chen HL, Copperman AB, Grunfeld L, Sandler B, Bustillo M, Gordon JW. Failed fertilization in vitro: second day micromanipulation of oocytes versus reinsemination. Fertil Steril. 1995;63:1337–40.
5. Nagy ZP, Joris H, Liu J, Staessen C, Devroey P, Van Steirteghem AC. Intracytoplasmic single sperm injection of 1-day-old unfertilized human oocytes. Hum Reprod. 1993;8:2180–4.
6. Chen C, Kattera S. Rescue ICSI of oocytes that failed to extrude the second polar body 6 h post-insemination in conventional IVF. Hum Reprod. 2003;18:2118–21.
7. Zhu L, Xi Q, Nie R, Chen W, Zhang H, Li Y. Rescue intracytoplasmic sperm injection: a prospective randomized study. J Reprod Med. 2011;56:410–4.
8. Jin H, Shu Y, Dai S, Peng Z, Shi S, Sun Y. The value of second polar body detection 4 hours after insemination and early rescue ICSI in preventing complete fertilisation failure in patients with borderline semen. Reprod Fertil Dev. 2014;26:346–50.
9. Liu W, Liu J, Zhang X, Han W, Xiong S, Huang G. Short co-incubation of gametes combined with early rescue ICSI: an optimal strategy for complete fertilization failure after IVF. Hum Fertil (Camb). 2014;17:50–5.
10. Matsunaga R, Morita H, Hasegawa R, Isobe K, Miura M, Kobayashi Y, et al. Identification of the indicators for rescue ICSI: the efficacy oftime-lapse imaging for the signs of fertilization inivf. Fertil Steril. 2019;112:e277.
11. Shibahara T, Fukasaku Y, Miyazaki N, Kawato H, Minoura H. Usefulness of expanding the indications of early rescue intracytoplasmic sperm injection. Reprod Med Biol. 2022;21:e12432.

12. Cao S, Wu X, Zhao C, Zhou L, Zhang J, Ling X. Determining the need for rescue intracytoplasmic sperm injection in partial fertilisation failure during a conventional IVF cycle. Andrologia. 2016;48:1138–44.
13. Abiodun OI, Kiru MU, Jantan A, Omolara AE, Dada KV, Umar AM, et al. Comprehensive review of artificial neural network applications to pattern recognition. IEEE Access. 2019;7:158820–158846.
14. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. NIPS. 2017:3149–57. Available from: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
15. Li L, Cui X, Yang J, Wu X, Zhao G. Using feature optimization and LightGBM algorithm to predict the clinical pregnancy outcomes after in vitro fertilization. Front Endocrinol (Lausanne). 2023;14:1305473.
16. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA. 2016; pp.1135-44. Available from: https://dl.acm.org/doi/10.1145/2939672.2939778
17. Kanda Y. Investigation of the freely available easy-to-use software "EZR" for medical statistics. Bone Marrow Transplant. 2013;48:452–8.
18. Geng J, Cai J, Ouyang L, Liu L, Liu Z, Ma C, et al. Indications affect neonatal outcomes following early rescue ICSI: a retrospective study. J Assist Reprod Genet. 2024;41:661–72.
19. Shalom-paz E, Alshalati J, Shehata F, Jimenez L, Son W-Y, Holzer H, et al. Clinical and economic analysis of rescue intracytoplasmic sperm injection cycles. Gynecol Endocrinol. 2011;27:993–6.
20. Fang Q, Jiang X, Bai S, Xu B, Zong L, Qi M, et al. Safety of early cumulus cell removal combined with early rescue ICSI in the prevention of fertilization failure. Reprod Biomed Online. 2023;47:103214.
21. Zeng J, Yao Z, Zhang Y, Tian F, Liao T, Wu L, et al. Fertilization and neonatal outcomes after early rescue intracytoplasmic sperm injection: a retrospective analysis of 16,769 patients. Arch Gynecol Obstet. 2022;306:249–58.
22. Dai S-J, Qiao Y-H, Jin H-X, Xin Z-M, Su Y-C, Sun Y-P, et al. Effect of coincubation time of sperm-oocytes on fertilization, embryonic development, and subsequent pregnancy outcome. Syst Biol Reprod Med. 2012;58:348–53.
23. Coticchio G, Mignini RM, Novara PV, Lain M, De PE, Turchi D, et al. Focused time-lapse analysis reveals novel aspects of human fertilization and suggests new parameters of embryo viability. Hum Reprod. 2018;33(1):23–31. https://doi.org/10.1093/humrep/dex344
24. Liang J. Image classification based on RESNET. J Phys Conf Ser. 2020;1634:012110.
25. Bhagat M, Bakariya B. A comprehensive review of cross-validation techniques in machine learning. Int J Sci Technol. 2025;16(1). Available from: https://www.ijsat.org/papers/2025/1/1305.pdf
26. Brusentsev EY, Mokrousova VI, Igonina TN, Rozhkova IN, Amstislavsky SY. Role of lipid droplets in the development of oocytes and preimplantation embryos in mammals. Russian Journal of Developmental Biology. 2019;50:230–7.
27. Terada Y, Morito Y, Tachibana M, Morita J, Nakamura S-I, Murakami T, et al. Cytoskeletal dynamics during mammalian gametegenesis and fertilization: implications for human reproduction. Reprod Med Biol. 2005;4:179–87.
28. Ibayashi M, Aizawa R, Mitsui J, Tsukamoto S. Homeostatic regulation of lipid droplet content in mammalian oocytes and embryos. J Reprod Fertil. 2021;162:R99–R109.
29. Van Blerkom J, Davis P, Merriam J, Sinclair J. Nuclear and cytoplasmic dynamics of sperm penetration, pronuclear formation and microtubule organization during fertilization and early preimplantation development in the human. Hum Reprod Update. 1995;1:429–61.
30. Morales DA, Bengoetxea E, Larrañaga P, García M, Franco Y, Fresnada M, et al. Bayesian classification for the selection of in vitro human embryos using morphological and clinical data. Comput Methods Programs Biomed. 2008;90:104–16.
31. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv [cs.LG]. 2020; arXiv:2010.16061. Available from: http://arxiv.org/abs/2010.16061
32. Takahashi K, Yamamoto K, Kuchiba A, Koyama T. Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores. Appl Intell. 2022;52:4961–72.
33. DeVries Z, Locke E, Hoda M, Moravek D, Phan K, Stratton A, et al. Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability. Spine J. 2021;21:1135–42.
34. Guo Y, Liu W, Wang Y, Pan J, Liang S, Ruan J, et al. Polarization microscopy imaging for the identification of unfertilized oocytes after short-term insemination. Fertil Steril. 2017;108:78–83.
35. Eichenlaub-Ritter U, Shen Y, Tinneberg H-R. Manipulation of the oocyte: possible damage to the spindle apparatus. Reprod Biomed Online. 2002;5:117–24.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Saito M, Haraguchi H, Nakajima I, Fukuda S, Zhu C, Masuya N, et al. A machine learning model for predicting fertilization following short-term insemination using embryo images. Reprod Med Biol. 2025;24:e12649. https://doi.org/10.1002/rmb2.12649