Title

Innovative AI models for clinical decision-making: predicting blastocyst formation and quality from time-lapse embryo images up to embryonic day 3

Running title

AI-Based Blastocyst Formation Prediction

Authors

A Yanai, MD[1], A Horie, MD, PhD[1, 2], A Sakurai, MD[1], S Imakita, MD[1], M Nakamura, MD[1], A Ikeda, MD[1], S Shitanaka, MD[1], T Ohara, MD[1], B Nakakita, MD[1], A Ueda, MD[1], Y Kitawaki, MD[1], Y Sagae, MD, PhD[1], A Okunomiya, MD, PhD[1], M Mandai, MD, PhD[1]

[1] Department of Gynecology and Obstetrics, Kyoto University Graduate School of Medicine, Kyoto, Japan

[2] Department of Obstetrics & Gynecology, Medical Research Institute KITANO HOSPITAL, PIIF Tazuke-kofukai, Osaka, Japan

Corresponding author:

Akihito Horie, MD, PhD,

ORCID iD: 0000-0002-0801-5835

Department of Gynecology and Obstetrics,

Kyoto University Graduate School of Medicine,

54 Shogoin Kawahara-cho, Sakyo, Kyoto 606-8507, Japan

Phone: 81-75-751-3269

Email: a_horie@kuhp.kyoto-u.ac.jp

Abstract

Accurate embryo assessment on embryonic day 3 of assisted reproductive technology (ART) is crucial for deciding whether to continue the culture until day 5 (blastocyst stage) or opt for earlier transfer or cryopreservation. Prolonged culture often improves pregnancy outcomes in patients with multiple high-quality embryos, but may offer limited benefits for older patients or those with few available embryos. To address this clinical challenge, analyzing embryo quality in early stages by artificial intelligence (AI) can be useful. We retrospectively analyzed images of 7,111 two-pronuclear embryos (Veeck grade ≤3) using four different time-lapse incubators.

We fine-tuned ImageNet-1k-pretrained NASNet-A Large to automatically classify each time-lapse image into 17 morphological categories, including cell stages and Veeck grades 1–3. This model achieved 95% cell-stage accuracy on the test set. We combined these annotations with age at egg retrieval in a gradient boosting framework (XGBoost) to predict blastocyst formation, good blastocysts, and poor blastocyst + arrested embryos (PBAE). The ROC AUCs were 0.87, 0.88, and 0.87 for blastocyst formation, good blastocysts, and PBAE, respectively, indicating good predictive performance for day 3 embryo assessment. Notably, the PBAE model reached a precision-recall AUC of 0.90, accurately identifying embryos unlikely to benefit from extended culture.

This revolutionary AI prediction model could ensure transparency and addresses the "black box" limitation often associated with AI. By integrating a high-accuracy auto-annotation pipeline with interpretable AI (via SHapley Additive exPlanations), our device-independent approach supports appropriate embryo-specific decisions, potentially reducing unnecessary culture, optimizing workflows, and improving clinical outcomes in ART.

Keywords

Artificial intelligence, blastocyst, prediction model, machine learning, time-lapse, assisted reproductive technology, reproductive medicine

Capsule: Artificial intelligence models accurately predicted blastocyst formation and quality using time-lapse images and age on embryonic day 3, supporting clinical decision-making regarding blastocyst culture or early embryo transfer.

Highlights:

- Time-lapse (24–64 hpi) enables accurate day 3-based blastocyst prediction (AUC=0.87).
- Poor blastocyst + arrested embryo model offers PR AUC=0.90 for early decisions.
- Robust across four incubator types and age groups, demonstrating broad utility.
- SHAP analysis clarifies how morphological features drive AI predictions.
- Embryo-specific decisions reduce unnecessary culture and improve patient outcomes.

Introduction

Assisted reproductive technology (ART) plays an important role in infertility treatment. In 2021, 9% of all births in Japan resulted from ART [1]. Within ART, there are two main timings for embryo transfer: cleavage-stage transfer (on the embryonic day 2 or 3) and blastocyst-stage transfer (on the embryonic day 5 or 6). In patients with a good prognosis, blastocyst-stage transfer is usually preferred owing to higher pregnancy and birth rates per transfer [2,3]. According to the Japanese National Registry of ART [1,4], 73% of transfer cycles in 2021 were blastocyst-stage transfers, and the pregnancy rate per embryo transfer was double that of cleavage-stage transfers (41% vs. 20%, respectively). Conversely, in patients with poor prognoses and few available embryos (e.g., over 40 years of age), prolonged culture sometimes fails to produce transferable blastocysts. Several reports [5–7] showed that extending culture to the blastocyst stage did not improve implantation or cumulative pregnancy rates in certain scenarios. Additionally, one report [8] pointed out that the live birth rate of cleavage-stage embryo transfer was double that of blastocyst culture in patients with only one available embryo on embryonic day 3. Thus, cleavage-stage embryo transfer remains an important option for patients with a poor prognosis, such as older patients. In Japan, where donor eggs are difficult to obtain, 47% of patients aged >43 years undergo cleavage–stage transfer [1].

In previous studies [2], the choice between early embryo freezing/transfer and blastocyst transfer was made on a patient-by-patient basis rather than an embryo-by-embryo basis. Embryo-by-embryo optimal selection requires accurate prediction of morphological quality on embryonic day 5 or 6 by embryonic day 3. Although one report [9] showed that static morphological assessment on embryonic day 3 could predict blastocyst formation, the

87 receiver operating characteristic area under the curve (ROC AUC) of 0.78 was considered
88 insufficient for reliable clinical decision-making.

89 Since 2008, time-lapse embryo culture [10] has enabled the evaluation of morphokinetic
90 parameters, leading to multiple studies reporting predictions of blastocyst formation
91 [11,12,21,13–20] and good blastocyst formation [18,20,22–24] using these parameters.
92 While day 4 predictions of blastocyst formation achieved validated AUCs as high as 0.85 [19],
93 clinically useful predictions at day 3 remained lower, with AUCs of 0.73–0.81 [17,21].

94 Recent advances have included artificial intelligence (AI) analysis of time-lapse images [25–
95 29], with one reported AI-based model achieving an AUC of 0.82 for predicting blastocyst
96 formation on embryonic day 3 [29]. However, previously reported predictive models
97 [9,17,21,29] face several limitations. First, the prediction accuracy was insufficient for
98 reliable clinical decisions regarding whether extended blastocyst culture or early embryo
99 transfer/freezing on embryonic day 3. Next, the time-lapse parameters required laborious
100 manual annotation and exhibited interrater variability [30]. Last, AI prediction models often
101 lacked interpretability (“black box” issue) and were highly device-dependent, as they were
102 trained using images from only one type of time-lapse incubator [25–29].

103 This study aimed to develop an accurate AI blastocyst prediction model on embryonic day 3
104 with high versatility and interpretability for clinical use. To achieve this objective, we
105 analyzed time-lapse images acquired from multiple time-lapse incubators at multiple
106 facilities, constructed models from human-interpretable features, and visualized the AI
107 decision basis using SHapley Additive exPlanations (SHAP) [31,32].

108

Materials and Methods

Overview of the Blastocyst Prediction using AI

An overview is shown in Figure 1. We created a dataset for the blastocyst prediction model (Figure 1A). We fine-tuned the ImageNet-1k pre-trained image analysis AI (NASNet-A Large [33]), using a 17-class embryo image classification dataset to build an AI auto-annotator (Figure 1B). AI extracted features from time-lapse images, and the age at oocyte retrieval was analyzed using a gradient boosting method (XGBoost [34]) to build each prediction model for blastocyst formation, good blastocyst formation, and poor blastocyst + arrested embryos (Figure 1E). Our AI blastocyst prediction model predicted embryonic morphology on embryonic day 5 or 6 based on embryo culture time-lapse images and age at oocyte retrieval.

Fine-Tuning the AI Auto-Annotator

A 17-class embryo image classification dataset (Figure 2A) was created from 23,852 embryo images with 17-class supervised labels, including 1-cell stage with or without pronuclei (PN) and 2-, 3-, 4-, 5–7-, and 8-cell stages with Veeck grades 1/2/3 [35]. The ImageNet-1k pre-trained image AI, NASNet-A Large [33], was fine-tuned on this dataset to create an AI auto-annotator (Supplementary Text 1). Time-lapse images were analyzed using the AI auto-annotator and classified into 17-class morphological labels. The proportions of each cell stage and Veeck grade were calculated in each embryo as features. The Veeck grade was replaced by the mode frequency (or a better grade if the mode frequency was the same) at each cell stage for each embryo.

Data Collection and Preparation for the Blastocyst Prediction Dataset

Embryo culture time-lapse images and clinical data recorded in daily practice were retrospectively collected from Kyoto University Hospital and its cooperating facilities from May 2018 to December 2022 (Supplementary Text 1 and Figure 1C).

This study included 2PN embryos with cleavage that were time-lapse cultured, with at least embryonic day 5, single embryo per well. Degenerated and poor embryos with Veeck grades 4 and 5 were excluded. Time-lapse images were collected every 15 min at 24–64 hours post-insemination (hpi). Time-lapse images were collected up to 64 hpi because most day 3 embryo transfer/freezing were performed after 64 hpi. Time-lapse videos were preprocessed and converted to JPG images of 331 × 331 pixels. Embryos with more than 5% image loss were excluded due to missing data, debris, misplacement or significant preprocessing errors. (Supplementary Text 1).

The blastocyst prediction dataset was created by combining the time-lapse images at each time point with the morphological evaluation on embryonic day 5 or 6, age at oocyte retrieval, and facility.

Overall, 8,043 fertilized embryos cultured in four types of time-lapse incubators in 2,792 oocyte retrieval cycles at four facilities were included. After exclusions, 7,111 embryos remained. The dataset was randomly split into 80% for training and cross-validation and 20% (1,422 embryos) for testing (Figure 1A).

Blastocysts were evaluated for Gardner stage and grade; blastocyst formation was defined as Gardner stage ≥3, and arrested embryos were defined as others. Good blastocysts were defined as Gardner grades AA, AB, BA, and BB; poor blastocysts as CC; and others as moderate blastocysts. The morphological evaluation is shown in Figure 1D.

153 **Building the AI Blastocyst Prediction Models**

154 The AI auto-annotator was used to classify the time-lapse images at each time point into
155 morphological labels, and the proportions of cell stages and Veeck grades were calculated.
156 Time-series labels, proportions, and ages at oocyte retrieval were analyzed using XGBoost
157 and models were built from the training and cross-validation dataset (Supplementary Text
158 1). The targets for prediction were blastocyst formation, good blastocyst formation, and
159 poor blastocyst + arrested embryos. The models were five-fold cross-validated and tested on
160 a test dataset. The endpoint was ROC AUC and the AUC for precision-recall (PR). For clinical
161 decision-making, a target PR AUC of poor blastocyst + arrested embryos prediction was set
162 at 0.90. Subgroup analyses were performed by facility and age group (<35, 35–39, and ≥40
163 years).

164 **Blastocyst Prediction Models from Single or Few Embryos Images**

165 To examine the superiority of time-lapse image analysis over static image analysis,
166 prediction models were created from a single image at 24, 48, or 64 hpi, or three images at
167 24, 48, and 64 hpi, assuming a once-a-day observation, and ROC AUC values were compared.

168 In the first type of model, the AI auto-annotator classified static images into morphological
169 annotation labels, and XGBoost built prediction models from single or three labels and age.
170 For the next type of model, the fine-tuned NASNet-A Large was trained on single images
171 (grayscale, 331 × 331 pixels) at 24, 48, and 64 hpi in the training and cross-validation dataset
172 to create prediction models. Additionally, because NASNet-A Large is a color image analysis
173 AI that allows up to three channels of input for each image, we created merged images using

24-, 48-, and 64-hpi images for each channel and trained NASNet-A Large on the merged images to create the prediction models.

SHAP-Based Model Interpretation

In our prediction models, the AI auto-annotator classified embryo images into human-recognizable morphological labels, and all features inputted to XGBoost were interpretable. However, because XGBoost has dozens to hundreds of decision trees, it is difficult to directly interpret the basis for the prediction models. Therefore, we used SHAP [31,32] to analyze the contribution and predictive direction of each feature in the test dataset. Because SHAP values are positive and negative for positive and negative predictions, respectively, we calculated the absolute mean of SHAP values for each feature and extracted the highly contributing features. The correlation between the value and SHAP value of each contributing feature was visualized and used to interpret the models.

Statistical Analyses

The model performance was evaluated using ROC and PR analyses and AUCs. The analyses were performed using numpy 1.21.6, scipy 1.7.1, and sklearn 1.0.2 on Python 3.7.15. In the comparison of blastocyst prediction models from time-lapse images and those from single or few embryo images, ROC AUCs were compared, and p-values were calculated using the DeLong test. The DeLong test was performed using the fast DeLong method in Python (https://github.com/yandexdataschool/roc_comparison) [36,37]. The significance level was

195 set at $p<0.05$; however, because of nine-fold tests, the significance level was set at $p<0.0055$
196 with Bonferroni correction.

197

198 **Ethical Considerations**

199 This study was conducted in accordance with the principles embodied in the Declaration of
200 Helsinki and was approved by the Ethics Committee of Kyoto University Graduate School and
201 Faculty of Medicine (approval number R3320). Informed consent was obtained in the form
202 of opt-out through the websites of facilities.

203

204 Results

205 **AI Auto-Annotator Performance**

206 We fine-tuned ImageNet-1k pre-trained image AI NASNet-A Large with the 17-class embryo
207 image classification dataset (Figures 1B and 2A). The training was early stopped at 30 epochs
208 (Figure 2B and C), and the test accuracy was 85%. The confusion matrix (Figure 2D) showed
209 that most misclassifications were due to errors in the Veeck grade, the cell stage accuracy
210 was excellent at 95% (Supplementary Figure 1).

211 The distribution of AI annotation labels at each time point in the blastocyst prediction
212 dataset showed the progression of cleavage from two to eight cells (Figure 2E). At 24 hpi, the
213 percentage of annotations misclassified as four or more cells owing to fragmentation or
214 other factors was small (2.0%).

215 Major time-lapse parameters were calculated from the AI annotation labels and compared
216 with those in previous reports [12,13,17,19,38] (Supplementary Table 1). AI-calculated t2
217 (time to the 2-cell stage) was significantly smaller in the blastocyst formation group than in
218 the arrested embryo group (median: 24.8 vs. 27.0 hpi, $p<0.001$) and was within the mean or
219 median values in previous reports.

220 **Blastocyst Formation Prediction Model**

221 Using XGBoost and recursive feature elimination, 88 features were selected for the
222 blastocyst formation prediction model (Supplementary Table 2). The model achieved a test
223 AUC of 0.87 and a five-fold cross-validation mean AUC of 0.86 (Figure 3A). With a Youden's
224 index [39] threshold of 0.49, the test accuracy reached 0.79. These results demonstrated the
225 model's robustness in cross-validation and testing.

226 **Good Blastocyst Formation Prediction Model**

227 The good blastocyst formation prediction model achieved a test ROC AUC of 0.88 using 156
228 features. Owing to the low proportion of good blastocysts (28%), the test PR AUC was
229 relatively lower than other models (Figure 3B).

230 **Poor Blastocyst + Arrested Embryo Prediction Model**

231 The poor blastocyst + arrested embryo prediction model (Figure 3C), using 76 features,
232 achieved a test AUC of 0.87. Its PR AUC reached 0.90, indicating superior positive predictive
233 accuracy compared to other models and meeting the target performance criteria. Designed
234 as a clinical tool, this model assists in selecting early embryo transfer or freezing on
235 embryonic day 3 instead of blastocyst culture. Adjusting the decision threshold further

236 personalizes the clinical application, enabling a focus on achieving a 90% positive predictive
237 value (PPV) at 60% sensitivity or an 80% PPV at 80% sensitivity (Figure 3D).

238 **Subgroup Analyses**

239 Subgroup analyses by facility for all models indicated no large differences between facilities,
240 and the blastocyst formation prediction model was able to predict with an accuracy of AUC
241 ≥0.83 for each facility (Figure 4A and Supplementary Figure 2A). Including age at oocyte
242 retrieval as a feature, subgroup analyses by age group showed small differences between
243 groups, and predictions were possible irrespective of the age group (Figure 4B and
244 Supplementary Figure 2B). These findings highlight the robustness of our models.

245 **Comparison of Time-Lapse and Limited-Image Models**

246 The AUC of the blastocyst formation prediction model built using XGBoost based on AI auto-
247 annotated labels from single or three images and age was 0.63–0.78; however, the AUC of
248 the model based on time-lapse image analysis was significantly higher, at 0.87 ($p<1.0\times10^{-5}$
249 for all comparisons; Figure 5A). The blastocyst formation prediction models based on image-
250 only analysis by NASNet-A Large had an AUC of 0.64–0.76, and the prediction model based
251 on time-lapse image analysis had a significantly higher AUC, at 0.87 ($p<1.0\times10^{-5}$ for all
252 comparisons) (Figure 5B). The models based on time-lapse image analysis were also
253 significantly better ($p<1.0\times10^{-5}$ for all comparisons) for good blastocyst formation prediction
254 and poor blastocyst + arrested embryo prediction (Supplementary Figure 3).

255 **Feature Contribution and SHAP Analysis**

256 The top 30 features for each model, ranked by absolute mean SHAP values (Figure 6A),
257 consistently included proportions (2-cell stage, 4-cell stage, Veeck grade 2, etc.) and age at

258 oocyte retrieval. While the time points of high contribution varied across models, a wide
259 range of features within the 24–64 hpi broadly contributed to the models.

260 The correlation between annotation labels and SHAP values (Figure 6B) at each time point in
261 the blastocyst formation prediction model indicated that whether each feature contributes
262 to a positive or negative prediction often varies at a single point because XGBoost deals with
263 the 17-class labels as discrete-value variables rather than categorical variables. The 5–7-cell
264 stage of grade 1/2 or 8-cell stage at 62.75 hpi, 2-cell stage or more at 29.75 hpi, and 4-cell
265 stage or less at 41.75 hpi were found to contribute to positive predictions. Focusing on 41.75
266 hpi, the 5-cell or more stage due to overestimation of poor-grade embryos contributed to
267 negative predictions, whereas the 3-cell or less stage, which was the most frequently
268 arrested stage, incorrectly contributed to positive predictions. In contrast, at 42.75 hpi, the
269 3-cell or less stage contributed appropriately to negative predictions, and the entire model
270 worked well.

271 The strong contribution of a very small proportion of 2-cell stage to negative predictions is
272 thought to represent abnormal direct cleavage.

273 Although our AI auto-annotator occasionally overestimated the cell stage in poor-grade
274 embryos, embryos with abnormally high proportion of 8-cell had negative SHAP values.
275 Thus, these blastocyst prediction models were able to accurately evaluate even poor-grade
276 embryos with morphological misclassification.

277

278 ## Discussion

279 **Main Findings**

In this study, 24–64-hpi time-lapse images recorded at various incubators and ages at oocyte retrieval from multiple centers were analyzed using AI to build models for predicting blastocyst morphology after the embryonic day 5. The AUCs for the blastocyst formation, good blastocyst formation, and poor blastocyst + arrested embryo prediction models were 0.87, 0.88, and 0.87, respectively. The positive predictive accuracy of the poor blastocyst + arrested embryo prediction model, which is envisioned as a clinical tool for deciding whether blastocyst culture should be continued, or early embryo freezing/transfer should be performed, achieved a PR AUC of 0.90, meeting the target accuracy.

Performance and Versatility of the AI Auto-Annotator

The newly developed AI auto-annotator offers several advantages over previously published models. In particular, it is compatible with various types of incubators, attains the highest level of cell stage classification accuracy without labor-intensive manual annotation, and incorporates additional qualitative evaluations, including pronuclearity and Veeck grade. Our cell stage accuracy of 95% compares favorably with previous studies [40–44]. This performance was particularly remarkable considering that we included fragmented embryos (Veeck grade 3) and more difficult-to-classify stages such as 3- and 5-7-cell. A previous study reported a slightly higher accuracy [41], but their model included fewer classification categories and a limited set of embryos, which may not reflect real-world diversity. In contrast, our approach includes a broader range of embryo states, including those with fragmentation, making it more generalizable to clinical practice.

Accuracy and Generalizability of the Prediction Models

301 We implemented versatile image pre-processing and classified time-lapse images at each
302 time point using the highly accurate AI auto-annotator, enabling detailed analysis of stage
303 and grade, calculation of parameters into robust prediction models.

304 Whereas prior day 3 blastocyst formation prediction models achieved validated ROC AUCs
305 ranging from 0.73 to 0.82 [9,17,21,29], our model demonstrated superior performance with
306 an AUC of 0.87. Similarly, our good blastocyst formation model achieved an ROC AUC of
307 0.88, improving on reported unvalidated AUCs of 0.73–0.82 [22–24] . Moreover, we
308 successfully created a poor blastocyst + arrested embryo model with a PR AUC of 0.90,
309 indicating its utility for identifying embryos unlikely to develop further. This prediction can
310 guide clinicians in deciding whether to proceed with day 3 transfer or freezing the embryo
311 rather than risk further culture.

312 Most previous models relied on time-lapse parameters or images captured within 24 hpi
313 [11,12,22–24,29,13–18,20,21]. However, our study included only images of the embryos
314 from 24 to 64 hpi, which are routinely recorded at most facilities, ensuring broader
315 applicability while maintaining superior accuracy.

316 In contrast to previous AI models, which were device-dependent [25–27,29], our study
317 analyzed images from four types of incubators across four facilities. By using a highly
318 adaptable pre-processing method and a versatile AI auto-annotator, our models effectively
319 managed variations such as aspect ratios, embryo area percentages, imaging contrasts,
320 resolutions, and well geometries.

321 Data were collected irrespective of oocyte retrieval age, insemination method, or culture
322 medium, creating heterogeneity across cases. Nonetheless, subgroup analyses showed
323 minimal differences in AUC across facilities and age groups, confirming the model's

generalizability. These findings suggest that our models are versatile, robust, and suitable for diverse clinical settings.

Interpretability of the prediction models

XGBoost outperformed Long Short-Term Memory models (data not shown) by analyzing features in parallel, effectively capturing temporal aspects of morphokinetic development from 24–64 hpi. SHAP analysis revealed how features, including human-recognizable parameters and embryo age, contributed to predictions. Importantly, localized misclassifications, such as overestimated cell stages in fragmented embryos, did not compromise overall model accuracy due to integration with more reliable features.

By restricting features to interpretable parameters, our approach ensures transparency and addresses the "black box" limitation often associated with AI. SHAP values allow clinicians to visualize individual embryo predictions, providing a clear understanding of the rationale behind the model's decisions. This interpretability fosters clinician confidence, accelerates regulatory approval, and supports the clinical application of AI tools in embryo assessment.

Limitations

Some limitations should be considered. First, we relied on the Gardner criteria [45,46] for blastocyst grading. While widely used, these criteria are subjective and prone to inter-rater variability [47]. To minimize domain shift, grading was conducted independently at each facility without standardization.

Second, embryos with missing records, misplacement, or poor-quality images were excluded, limiting absolute generalizability. However, this approach ensured a focus on clinically relevant embryos, enhancing practical applicability.

Finally, clinical heterogeneity in insemination methods, culture media, and patient demographics may have introduced variability. While our multicenter design mitigated these effects to some extent, further validation in prospective cohorts is warranted.

Despite these limitations, the study demonstrates the potential of AI models to advance embryo assessment in diverse clinical settings.

Conclusions

In this study, we developed an interpretable AI model capable of accurately predicting blastocyst morphology based on time-lapse images taken between 24 and 64 hpi and maternal age. The model's compatibility with various types of incubators ensures its broad applicability across diverse clinical settings. Notably, the poor blastocyst + arrested embryo prediction model offers significant clinical value, providing actionable insights for determining whether to continue blastocyst culture or perform early embryo freezing or transfer on day 3. By enabling embryo-specific decisions, this tool has the potential to optimize in vitro fertilization workflows, enhance individualized care, and improve patient outcomes.

Data availability: The AI code and required weights for this study are available at GitHub (https://github.com/KU-ObGy-ART/blastocyst-prediction/). Owing to the nature of this research, the participants of this study did not agree that their data, including time-lapse

images, should be shared publicly. Thus, images and clinical features of the fine-tuning and blastocyst prediction datasets are not available.

Acknowledgments

The authors thank Dr. Aisaku Fukuda at the IVF Osaka Clinic, Dr. Kimihiko Nakamura at the Amanogawa Ladies Clinic, and Dr. Masahide Shiotani and Noritoshi Enatsu at the Hanabusa Women's Clinic for generously providing the time-lapse images and clinical data.

Author's roles

A Yanai: Conceptualization, Methodology, Software, Investigation, Resources, Writing - Original Draft, Visualization, Formal analysis, Data Curation, Writing - Review & Editing. **A Horie:** Conceptualization, Methodology, Writing - Review & Editing, Project administration. **A Sakurai:** Investigation, Writing - Review & Editing. **S Imakita:** Investigation, Writing - Review & Editing. **M Nakamura:** Investigation, Writing - Review & Editing. **A Ikeda:** Investigation, Writing - Review & Editing. **S Shitanaka:** Validation, Writing - Review & Editing. **T Ohara:** Validation, Writing - Review & Editing. **B Nakakita:** Conceptualization, Methodology, Writing - Review & Editing. **A Ueda:** Methodology, Software, Resources, Writing - Review & Editing. **Y Kitawaki:** Investigation, Writing - Review & Editing. **Y Sagae:** Methodology, Funding acquisition, Writing - Review & Editing. **A Okunomiya:** Methodology, Resources, Writing - Review & Editing, **M Mandai:** Supervision, Writing - Review & Editing.

Funding

This work was supported by JSPS KAKENHI grant number JP23K08821.

Conflict of interest

None declared.

References

[1] Y. Katagiri, S.C. Jwa, A. Kuwahara, T. Iwasa, M. On, K. Kato, H. Kishi, Y. Kuwabara, F. Taniguchi, M. Harada, A. Iwase, Y. Osuga, Assisted reproductive technology in Japan: A summary report for 2021 by the Ethics Committee of the Japan Society of Obstetrics and Gynecology., Reprod Med Biol 23 (2024) e12552. https://doi.org/10.1002/rmb2.12552.

[2] D. Glujovsky, A.M. Quinteiro Retamar, C.R. Alvarez Sedo, A. Ciapponi, S. Cornelisse, D. Blake, Cleavage-stage versus blastocyst-stage embryo transfer in assisted reproductive technology., Cochrane Database Syst Rev 5 (2022) CD002118. https://doi.org/10.1002/14651858.CD002118.pub6.

[3] N.J. Cameron, S. Bhattacharya, D.J. McLernon, Cumulative live birth rates following blastocyst- versus cleavage-stage embryo transfer in the first complete cycle of IVF: A population-based retrospective cohort study, Hum Reprod 35 (2020) 2365–2374. https://doi.org/10.1093/humrep/deaa186.

[4] the Japan Society of Obstetrics and Gynecology ART Databook 2021, (n.d.). https://www.jsog.or.jp/activity/art/2021_JSOG-ART.pdf.

[5] V. Vlaisavljević, B. Kovačič, M. Reljič, V.G. Lovrec, M.Č. Sajko, Is there any benefit from the culture of a single oocyte to a blastocyst-stage embryo in unstimulated cycles?, Hum Reprod 16 (2001) 2379–2383. https://doi.org/10.1093/humrep/16.11.2379.

[6] B. Kovačič, V. Vlaisavljevi, M. Reljič, V. Gavri Lovrec, Clinical outcome of day 2 versus day 5 transfer in cycles with one or two developed embryos, Fertil Steril 77 (2002) 529–536. https://doi.org/10.1016/S0015-0282(01)03212-5.

[7] I. De Croo, R. Colman, P. De Sutter, D. Stoop, K. Tilleman, No difference in cumulative live birth rates between cleavage versus blastocyst transfer in patients with four or fewer zygotes: results from a retrospective study, Hum Reprod Open 2022 (2022) 1–9. https://doi.org/10.1093/hropen/hoac031.

[8] J.S. Xiao, M. Healey, A. Talmor, B. Vollenhoven, When only one embryo is available, is it better to transfer on Day 3 or to grow on?, Reprod Biomed Online 39 (2019) 916–923. https://doi.org/10.1016/j.rbmo.2019.08.003.

[9] X. Jiang, J. Cai, L. Liu, Z. Liu, W. Wang, J. Chen, C. Yang, J. Geng, C. Ma, J. Ren, Does conventional morphological evaluation still play a role in predicting blastocyst formation?, Reprod Biol Endocrinol 20 (2022) 1–12. https://doi.org/10.1186/s12958-022-00945-y.

[10] A.S. Leung, W.Y. Son, M.H. Dahan, Time-lapse imaging of embryos: current evidence supporting its use, Expert Rev Med Devices 13 (2016) 881–883. https://doi.org/10.1080/17434440.2016.1230015.

430 [11] C.C. Wong, K.E. Loewke, N.L. Bossert, B. Behr, C.J. De Jonge, T.M. Baer, R.A.R. Pera,
431 Non-invasive imaging of human embryos before embryonic genome activation
432 predicts development to the blastocyst stage, Nat Biotechnol 28 (2010) 1115–1121.
433 https://doi.org/10.1038/nbt.1686.

434 [12] M. Cruz, N. Garrido, J. Herrero, I. Pérez-Cano, M. Muñoz, M. Meseguer, Timing of cell
435 division in human cleavage-stage embryos is linked with blastocyst formation and
436 quality, Reprod Biomed Online 25 (2012) 371–381.
437 https://doi.org/10.1016/j.rbmo.2012.06.017.

438 [13] M. Dal Canto, G. Coticchio, M. Mignini Renzini, E. De Ponti, P.V. Novara, F.
439 Brambillasca, R. Comi, R. Fadini, Cleavage kinetics analysis of human embryos predicts
440 development to blastocyst and implantation, Reprod Biomed Online 25 (2012) 474–
441 480. https://doi.org/10.1016/j.rbmo.2012.07.016.

442 [14] S. Chamayou, P. Patrizio, G. Storaci, V. Tomaselli, C. Alecci, C. Ragolia, C. Crescenzo, A.
443 Guglielmino, The use of morphokinetic parameters to select all embryos with full
444 capacity to implant, J Assist Reprod Genet 30 (2013) 703–710.
445 https://doi.org/10.1007/s10815-013-9992-2.

446 [15] J. Conaghan, A.A. Chen, S.P. Willman, K. Ivani, P.E. Chenette, R. Boostanfar, V.L. Baker,
447 G.D. Adamson, M.E. Abusief, M. Gvakharia, K.E. Loewke, S. Shen, Improving embryo
448 selection using a computer-automated time-lapse image analysis test plus day 3
449 morphology: results from a prospective multicenter trial., Fertil Steril 100 (2013) 412–
450 9.e5. https://doi.org/10.1016/j.fertnstert.2013.04.021.

451 [16] K. Athayde Wirka, A.A. Chen, J. Conaghan, K. Ivani, M. Gvakharia, B. Behr, V. Suraj, L.
452 Tan, S. Shen, Atypical embryo phenotypes identified by time-lapse microscopy: High
453 prevalence and association with embryo development, Fertil Steril 101 (2014) 1637-
454 1648.e5. https://doi.org/10.1016/j.fertnstert.2014.02.050.

455 [17] R. Milewski, P. Kuć, A. Kuczyńska, B. Stankiewicz, K. Łukaszuk, W. Kuczyński, A
456 predictive model for blastocyst formation based on morphokinetic parameters in
457 time-lapse monitoring of embryo development, J Assist Reprod Genet 32 (2015) 571–
458 579. https://doi.org/10.1007/s10815-015-0440-3.

459 [18] S.T. Yang, J.X. Shi, F. Gong, S.P. Zhang, C.F. Lu, K. Tan, L.Z. Leng, M. Hao, H. He, Y.F. Gu,
460 G.X. Lu, G. Lin, Cleavage pattern predicts developmental potential of day 3 human
461 embryos produced by IVF, Reprod Biomed Online 30 (2015) 625–634.
462 https://doi.org/10.1016/j.rbmo.2015.02.008.

463 [19] Y. Motato, M.J. de los Santos, M.J. Escriba, B.A. Ruiz, J. Remohí, M. Meseguer,
464 Morphokinetic analysis and embryonic prediction for blastocyst formation through an
465 integrated time-lapse system., Fertil Steril 105 (2016) 376–84.e9.
466 https://doi.org/10.1016/j.fertnstert.2015.11.001.

467 [20] Y. Mizobe, M. Tokunaga, N. Oya, R. Iwakiri, N. Yoshida, Y. Sato, N. Onoue, Y. Ezono,
468 Synchrony of the first division as an index of the blastocyst formation rate during
469 embryonic development, Reprod Med Biol 17 (2018) 64–70.
470 https://doi.org/10.1002/rmb2.12070.

471 [21] N. Zaninovic, M. Nohales, Q. Zhan, Z.M.J. de los Santos, J. Sierra, Z. Rosenwaks, M.
472 Meseguer, A comparison of morphokinetic markers predicting blastocyst formation

473 and implantation potential from two large clinical data sets, J Assist Reprod Genet 36
474 (2019) 637–646. https://doi.org/10.1007/s10815-018-1396-x.

475 [22] K. Kirkegaard, U.S. Kesmodel, J.J. Hindkjær, H.J. Ingerslev, Time-lapse parameters as
476 predictors of blastocyst development and pregnancy outcome in embryos from good
477 prognosis patients: A prospective cohort study, Hum Reprod 28 (2013) 2643–2651.
478 https://doi.org/10.1093/humrep/det300.

479 [23] A. Storr, C.A. Venetis, S. Cooke, D. Susetio, S. Kilani, W. Ledger, Morphokinetic
480 parameters using time-lapse technology and day 5 embryo quality: a prospective
481 cohort study, J Assist Reprod Genet 32 (2015) 1151–1160.
482 https://doi.org/10.1007/s10815-015-0534-y.

483 [24] M. Cetinkaya, C. Pirkevi, H. Yelke, Y.K. Colakoglu, Z. Atayurt, S. Kahraman, Relative
484 kinetic expressions defining cleavage synchronicity are better predictors of blastocyst
485 formation and quality than absolute time points, J Assist Reprod Genet 32 (2015) 27–
486 35. https://doi.org/10.1007/s10815-014-0341-x.

487 [25] D. Tran, S. Cooke, P.J. Illingworth, D.K. Gardner, Deep learning as a predictive tool for
488 fetal heart pregnancy following time-lapse incubation and blastocyst transfer., Hum
489 Reprod 34 (2019) 1011–1018. https://doi.org/10.1093/humrep/dez064.

490 [26] C.I. Lee, Y.R. Su, C.H. Chen, T.A. Chang, E.E.S. Kuo, W.L. Zheng, W.T. Hsieh, C.C. Huang,
491 M.S. Lee, M. Liu, End-to-end deep learning for recognition of ploidy status using time-
492 lapse videos, J Assist Reprod Genet 38 (2021) 1655–1663.
493 https://doi.org/10.1007/s10815-021-02228-8.

494 [27] B. Huang, W. Tan, Z. Li, L. Jin, An artificial intelligence model (euploid prediction
495 algorithm) can predict embryo ploidy status based on time-lapse data, Reprod Biol
496 Endocrinol 19 (2021) 1–10. https://doi.org/10.1186/s12958-021-00864-4.

497 [28] J. Barnes, M. Brendel, V.R. Gao, S. Rajendran, J. Kim, Q. Li, J.E. Malmsten, J.T. Sierra, P.
498 Zisimopoulos, A. Sigaras, P. Khosravi, M. Meseguer, Q. Zhan, Z. Rosenwaks, O.
499 Elemento, N. Zaninovic, I. Hajirasouliha, A non-invasive artificial intelligence approach
500 for the prediction of human blastocyst ploidy: a retrospective model development and
501 validation study, Lancet Digit Heal 5 (2023) e28–e40. https://doi.org/10.1016/S2589-
502 7500(22)00213-8.

503 [29] Q. Liao, Q. Zhang, X. Feng, H. Huang, H. Xu, B. Tian, J. Liu, Q. Yu, N. Guo, Q. Liu, B.
504 Huang, D. Ma, J. Ai, S. Xu, K. Li, Development of deep learning algorithms for
505 predicting blastocyst formation and quality by time-lapse monitoring., Commun Biol 4
506 (2021) 415. https://doi.org/10.1038/s42003-021-01937-1.

507 [30] L. Sundvall, H.J. Ingerslev, U. Breth Knudsen, K. Kirkegaard, Inter- and intra-observer
508 variability of time-lapse annotations., Hum Reprod 28 (2013) 3215–3221.
509 https://doi.org/10.1093/humrep/det366.

510 [31] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv
511 Neural Inf Process Syst (2017) 4766–4775.

512 [32] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent Individualized Feature Attribution for
513 Tree Ensembles, (2018). http://arxiv.org/abs/1802.03888.

[33] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, Learning Transferable Architectures for Scalable Image Recognition, in: 2018 IEEE/CVF Conf Comput Vis Pattern Recognit, IEEE, 2018: pp. 8697–8710. https://doi.org/10.1109/CVPR.2018.00907.

[34] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, Proc ACM SIGKDD Int Conf Knowl Discov Data Min 13-17-Augu (2016) 785–794. https://doi.org/10.1145/2939672.2939785.

[35] L.L. Veeck, Oocyte Assessment and Biological Performance, Ann N Y Acad Sci 541 (1988) 259–274. https://doi.org/10.1111/j.1749-6632.1988.tb22263.x.

[36] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach., Biometrics 44 (1988) 837–45.

[37] X. Sun, W. Xu, Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves, IEEE Signal Process Lett 21 (2014) 1389–1393. https://doi.org/10.1109/LSP.2014.2337313.

[38] S. Canosa, L. Bergandi, C. Macrì, L. Charrier, C. Paschero, A. Carosso, N. Di Segni, F. Silvagno, G. Gennarelli, C. Benedetto, A. Revelli, Morphokinetic analysis of cleavage stage embryos and assessment of specific gene expression in cumulus cells independently predict human embryo development to expanded blastocyst: a preliminary study, J Assist Reprod Genet 37 (2020) 1409–1420. https://doi.org/10.1007/s10815-020-01806-6.

[39] W.J. Youden, Index for rating diagnostic tests., Cancer 3 (1950) 32–5. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

536 [40] A. Khan, S. Gould, M. Salzmann, Deep convolutional neural networks for human
537 embryonic cell counting, Lect Notes Comput Sci (Including Subser Lect Notes Artif
538 Intell Lect Notes Bioinformatics) 9913 LNCS (2016) 339–348.
539 https://doi.org/10.1007/978-3-319-46604-0_25.

540 [41] D. Dirvanauskas, R. Maskeliunas, V. Raudonis, R. Damasevicius, Embryo development
541 stage prediction algorithm for automated time lapse incubators, Comput Methods
542 Programs Biomed 177 (2019) 161–174. https://doi.org/10.1016/j.cmpb.2019.05.027.

543 [42] V. Raudonis, A. Paulauskaite-Taraseviciene, K. Sutiene, D. Jonaitis, Towards the
544 automation of early-stage human embryo development detection, Biomed Eng Online
545 18 (2019) 1–20. https://doi.org/10.1186/s12938-019-0738-y.

546 [43] B.D. Leahy, W.-D. Jang, H.Y. Yang, R. Struyven, D. Wei, Z. Sun, K.R. Lee, C. Royston, L.
547 Cam, Y. Kalma, F. Azem, D. Ben-Yosef, H. Pfister, D. Needleman, Automated
548 Measurements of Key Morphological Features of Human Embryos for IVF., Med Image
549 Comput Comput Assist Interv 12265 (2020) 25–35. https://doi.org/10.1007/978-3-
550 030-59722-1_3.

551 [44] G.B. Danardono, A. Erwin, J. Purnama, N. Handayani, A.A. Polim, A. Boediono, I. Sini, A
552 Homogeneous Ensemble of Robust Pre-defined Neural Network Enables Automated
553 Annotation of Human Embryo Morphokinetics, J Reprod Infertil 23 (2022) 250–256.
554 https://doi.org/10.18502/jri.v23i4.10809.

555 [45] D.K. Gardner, In-vitro culture of human blastocyst, in: Towar Reprod Certain Infertil
556 Genet beyond 1999, Parthenon Press, 1999: pp. 378–388.

557 [46] D.K. Gardner, M. Lane, J. Stevens, T. Schlenker, W.B. Schoolcraft, Blastocyst score
558 affects implantation and pregnancy outcome: Towards a single blastocyst transfer,
559 Fertil Steril 73 (2000) 1155–1158. https://doi.org/10.1016/S0015-0282(00)00518-5.

560 [47] A. Storr, C.A. Venetis, S. Cooke, S. Kilani, W. Ledger, Inter-observer and intra-observer
561 agreement between embryologists during selection of a single Day 5 embryo for
562 transfer: A multicenter study, Hum Reprod 32 (2017) 307–314.
563 https://doi.org/10.1093/humrep/dew330.

564

Figure legends

Figure 1. Overview of our blastocyst prediction AI. (**A**) Flowchart of the blastocyst prediction dataset. (**B**) Flowchart of the 17-class embryo image classification dataset and fine-tuning of AI auto-annotator. (**C**) Proportion of each facility in the blastocyst prediction dataset. (**D**) Proportion of embryo morphology on embryonic day 5 or 6 in the blastocyst prediction dataset. (**E**) Schema of the blastocyst prediction AI structure. AI = artificial intelligence, CV = cross-validation, n = number of embryos, Fac. = facility, SHAP = SHapley Additive exPlanations

Figure 2. Fine-tuning of the AI auto-annotator. (**A**) Labels and number of images in the 17-class embryo image classification dataset. (**B, C**) Accuracy and loss of fine-tuning. (**D**) Confusion matrix of fine-tuning (17-class image classification). (**E**) Proportion of AI annotation labels at each time point in the blastocyst prediction dataset. AI = artificial intelligence.

Figure 3. Performance of the prediction models. (**A, B, C**) ROC and precision-recall analyses of three prediction models. The blue lines indicate the results of five-fold validations, whereas the red line denotes the test result. (**D**) Precisions and recalls at each threshold in the poor blastocyst + arrested embryo prediction model. AUC = area under the curve, CV = cross-validation, ROC = receiver operating characteristic, PPV = positive predictive value

586 **Figure 4.** ROC subgroup analyses stratified by facility (**A**) and age group (**B**). AUC = area
587 under the curve, ROC = receiver operating characteristic, y.o. = years old

588

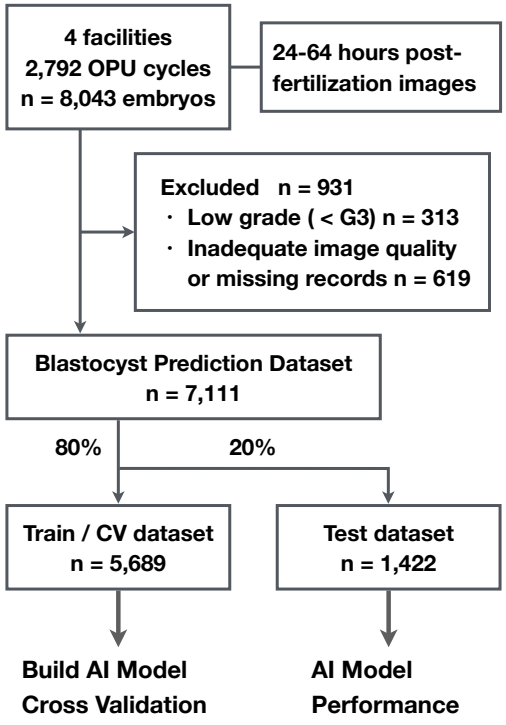
589 **Figure 5.** Blastocyst prediction models from single or few time-lapse images. (**A**) Comparison
590 of time-lapse-based model with one or three label-based models. (**B**) Comparison of time-
591 lapse-based model with one of three image-based models. AUC = area under the curve, hpi
592 = hours post-insemination

593

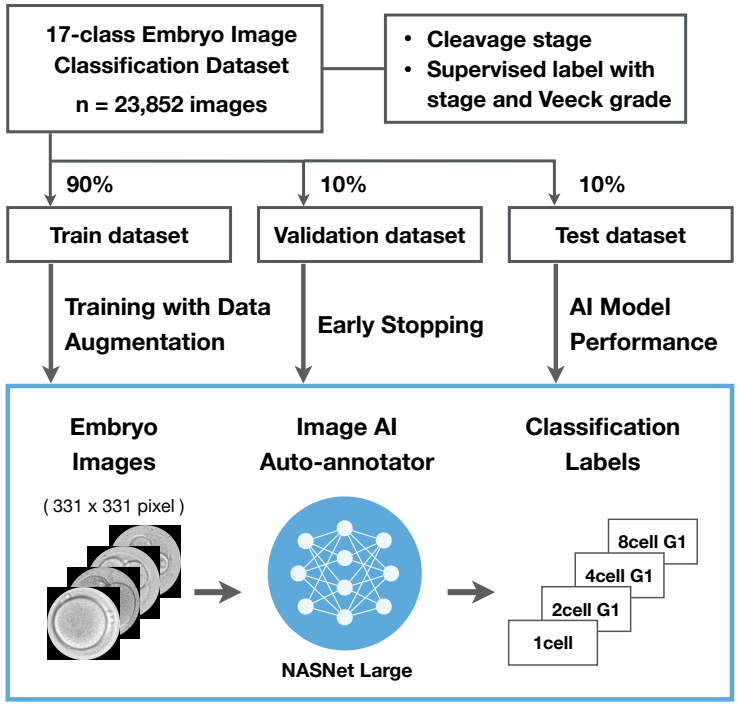
594 **Figure 6.** Visualization of model decision basis. (**A**) Top 30 contributing features based on
595 SHAP. (**B**) SHAP values of highly contributing features of the blastocyst formation prediction
596 model. SHAP values of highly contributing features of the good blastocyst formation and
597 poor blastocyst + arrested embryo prediction models are shown in Supplementary Figure 4.
598 SHAP = SHapley Additive exPlanations

599

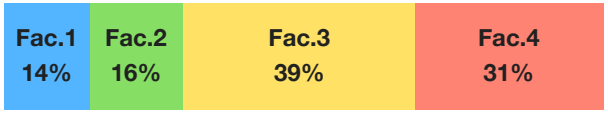
A. Blastocyst Prediction Dataset



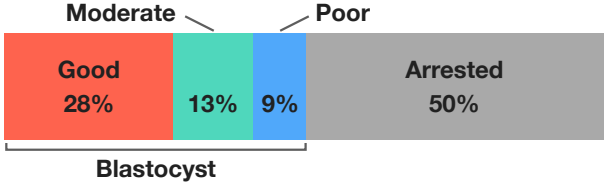
B. Fine-tuning of Auto-annotator



C. Proportion of Each Facility in the Blastocyst Prediction Dataset



D. Day5-6 Embryo Morphology of the Blastocyst Prediction Dataset



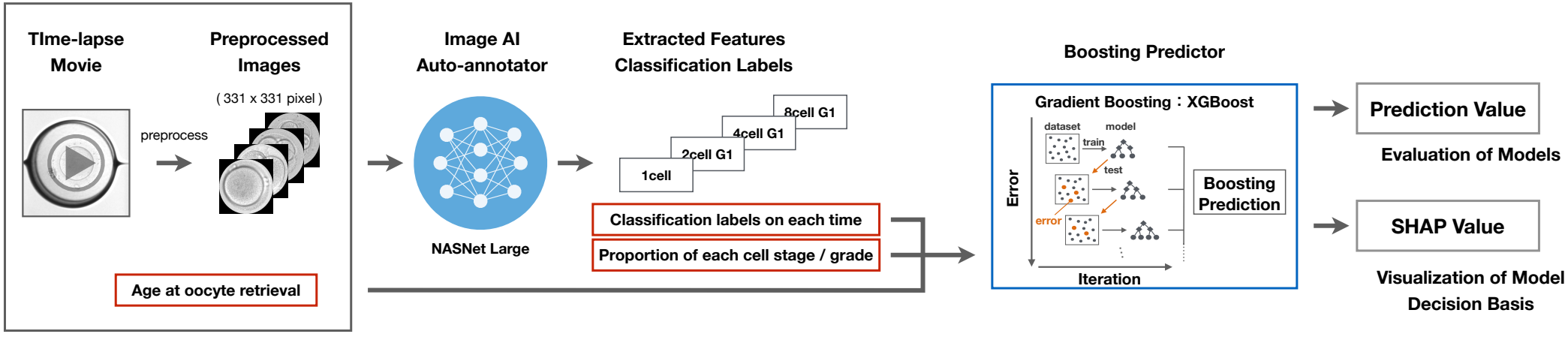
Classified by Gardner Grade

Good Blastocyst* : AA/AB/BA/BB

Moderate Blastocyst* : AC/CA/BC/CB

Poor Blastocyst*: CC * Gardner Stage ≥ 3

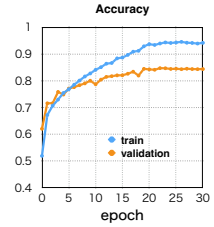
E. Blastocyst Prediction AI



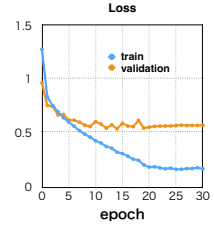
A.

Label	n
1cell	1,188
1cell PN	1,516
2cell G1	1,289
2cell G2	1,275
2cell G3	1,393
3cell G1	1,119
3cell G2	1,059
3cell G3	1,112
4cell G1	1,315
4cell G2	1,376
4cell G3	1,357
5-7cell G1	1,530
5-7cell G2	1,909
5-7cell G3	1,720
8-cell G1	1,537
8-cell G2	1,570
8-cell G3	1,587

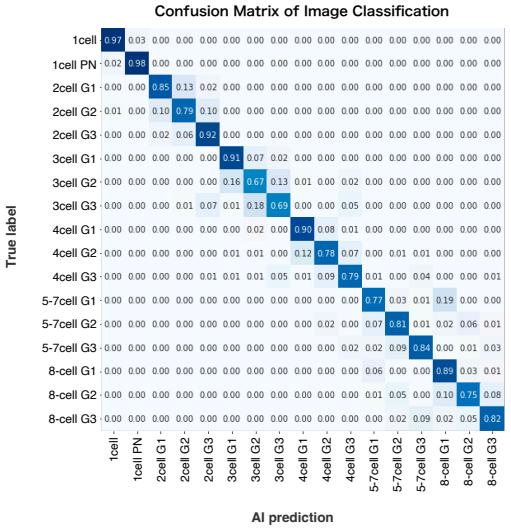
B.



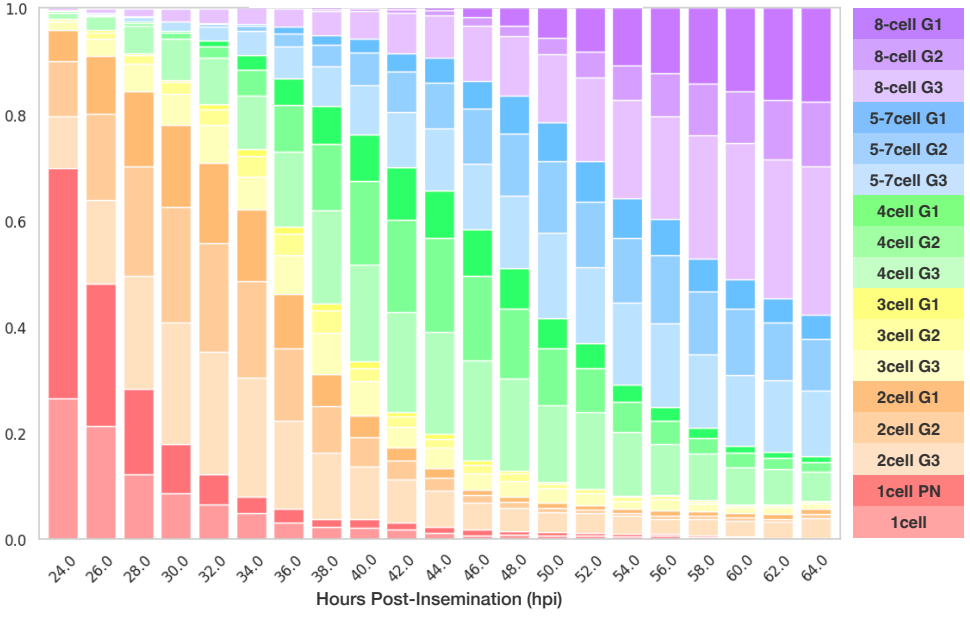
C.



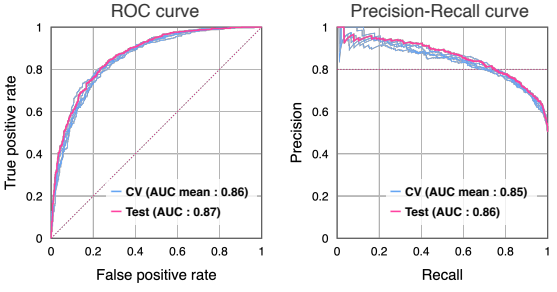
D.



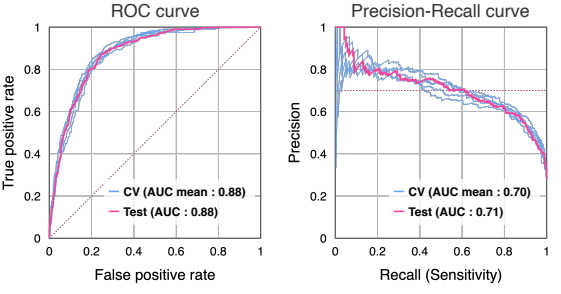
E. The proportion of AI annotation labels on each time in the blastocyst prediction dataset



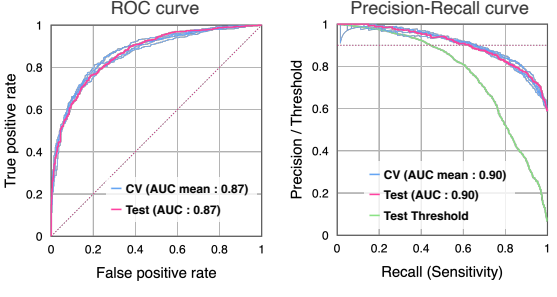
A. Blastocyst formation prediction model



B. Good blastocyst formation prediction model



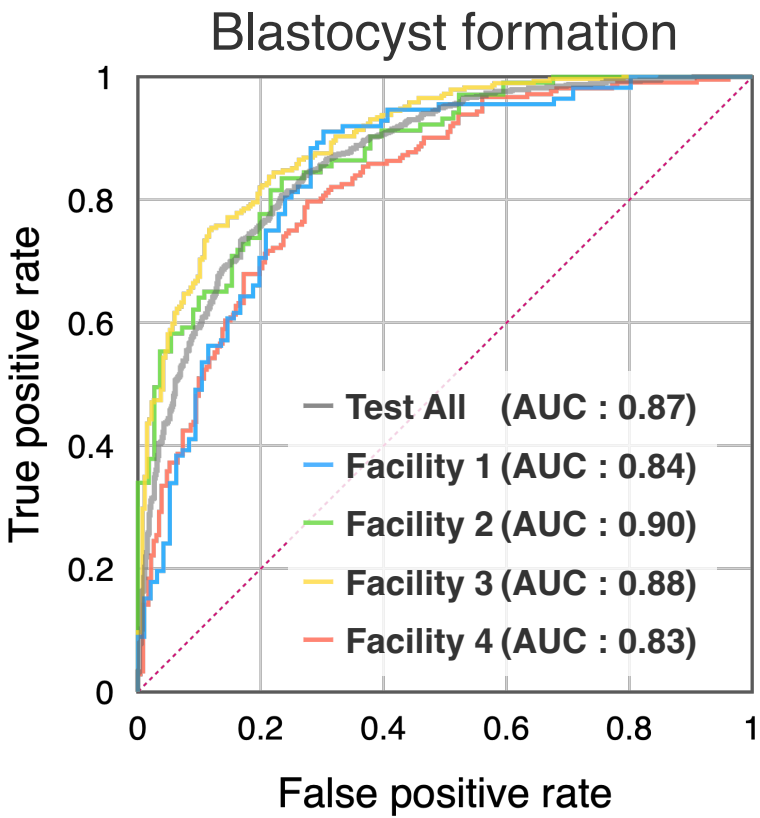
C. Poor blastocyst + arrested embryo prediction model



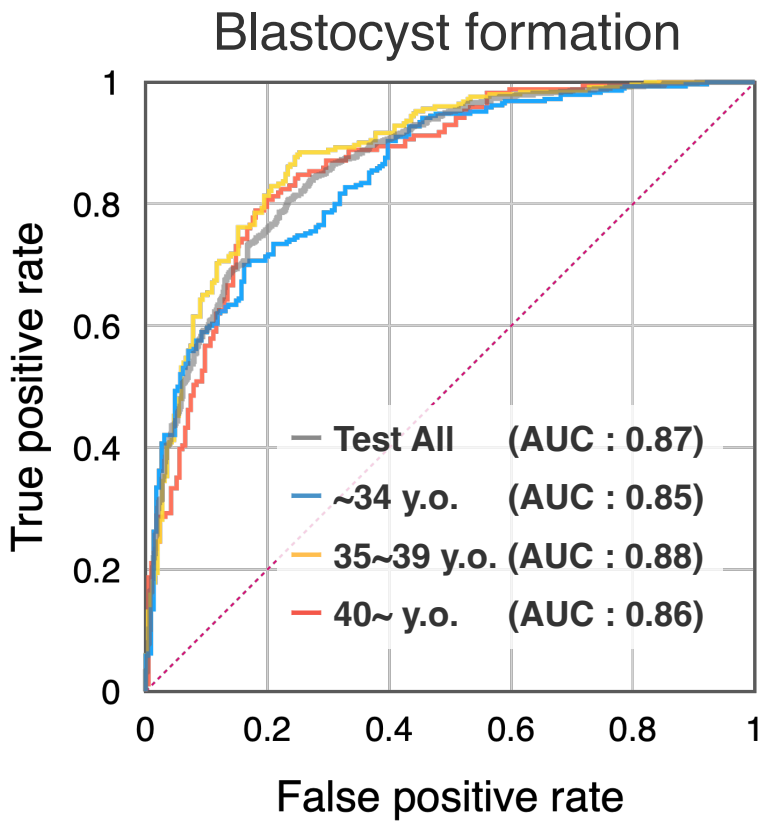
D. Poor blastocyst + arrested embryo prediction model

Threshold	Cross-validation mean		Test	
	Precision (PPV)	Recall (Sensitivity)	Precision (PPV)	Recall (Sensitivity)
0.9	0.95	0.45	0.95	0.44
0.8	0.92	0.60	0.90	0.61
0.7	0.89	0.69	0.87	0.70
0.6	0.85	0.76	0.85	0.76
0.5	0.83	0.81	0.82	0.81
0.4	0.79	0.87	0.79	0.87
0.3	0.74	0.92	0.75	0.92
0.2	0.68	0.96	0.69	0.97
Prior probability	0.59		0.59	

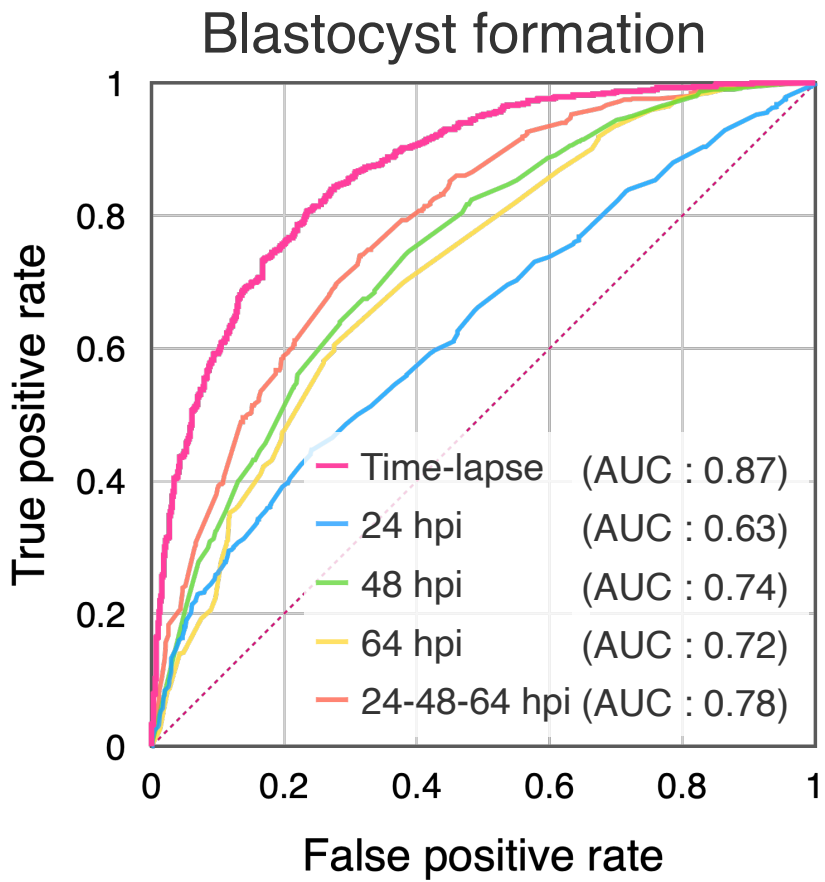
A. Subanalysis by facility



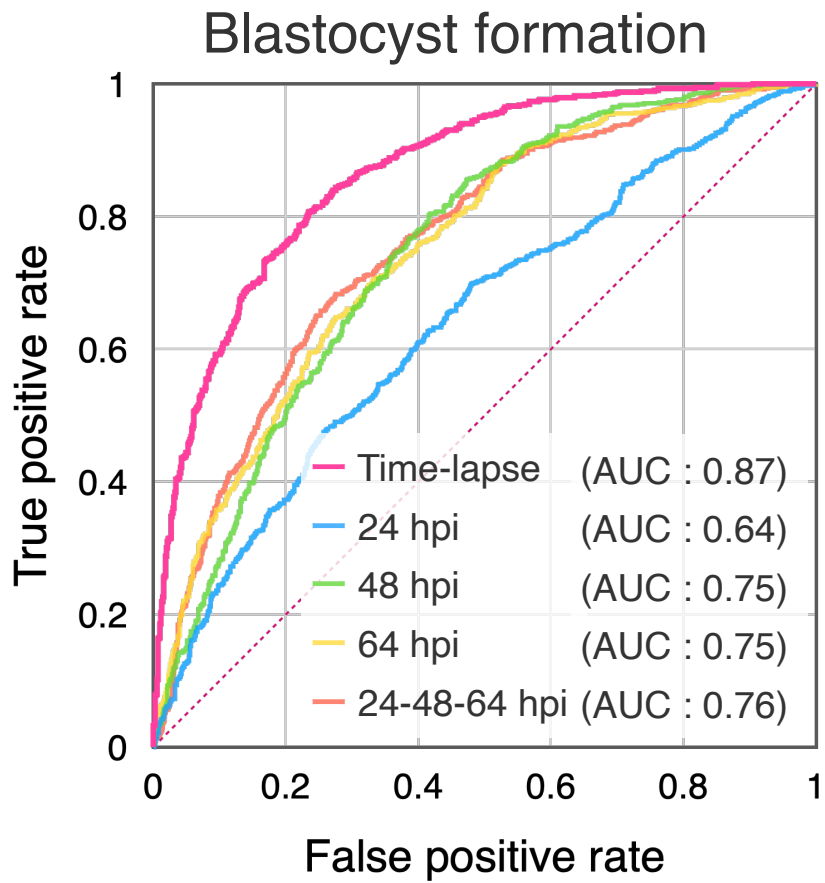
B. Subanalysis by age group



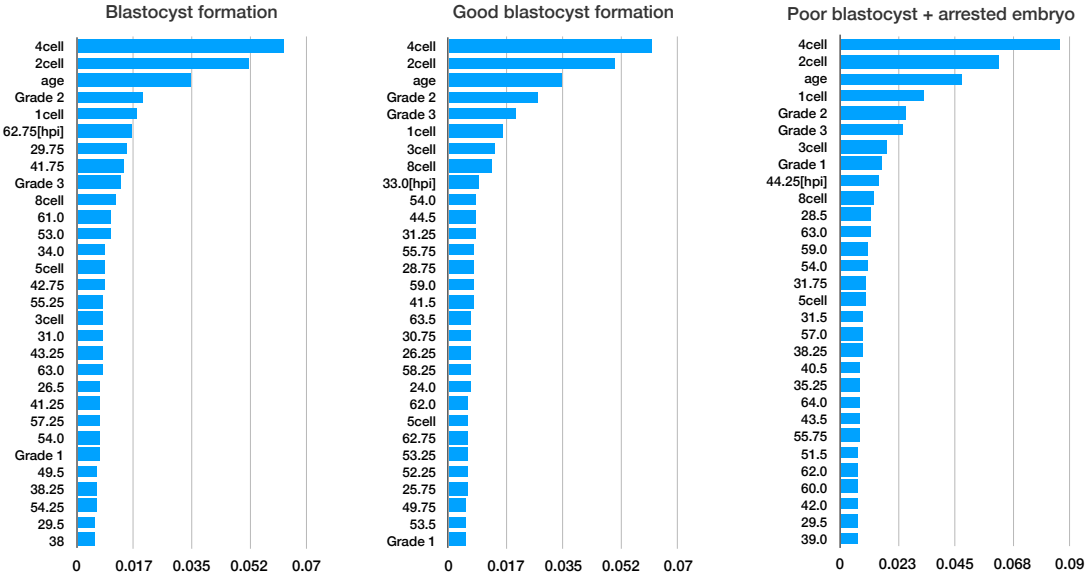
A. Annotation-based models



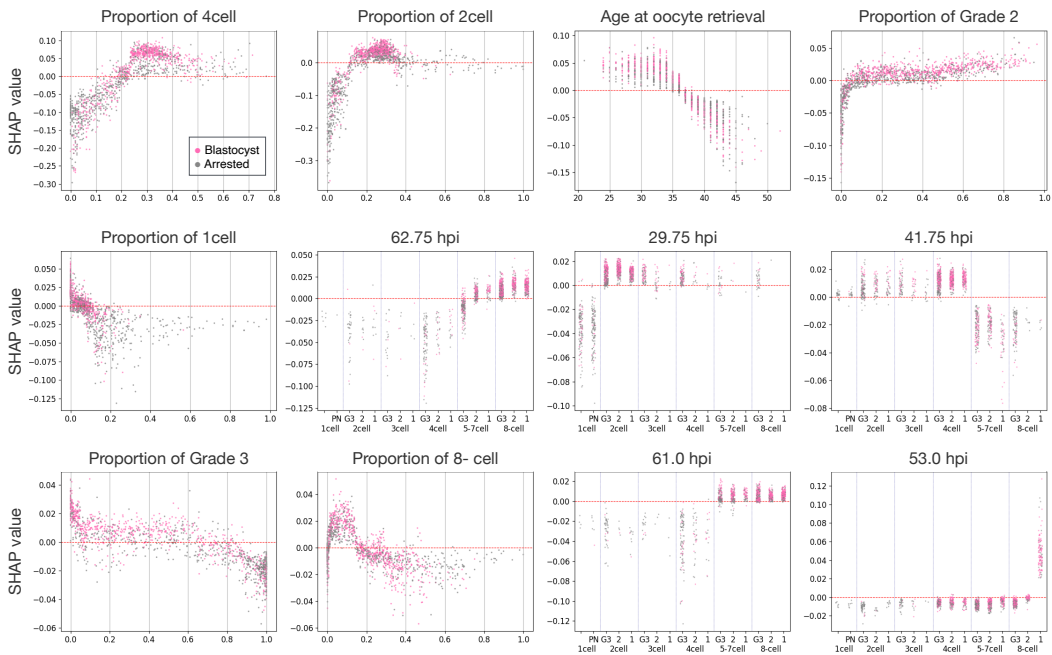
B. Image-based models

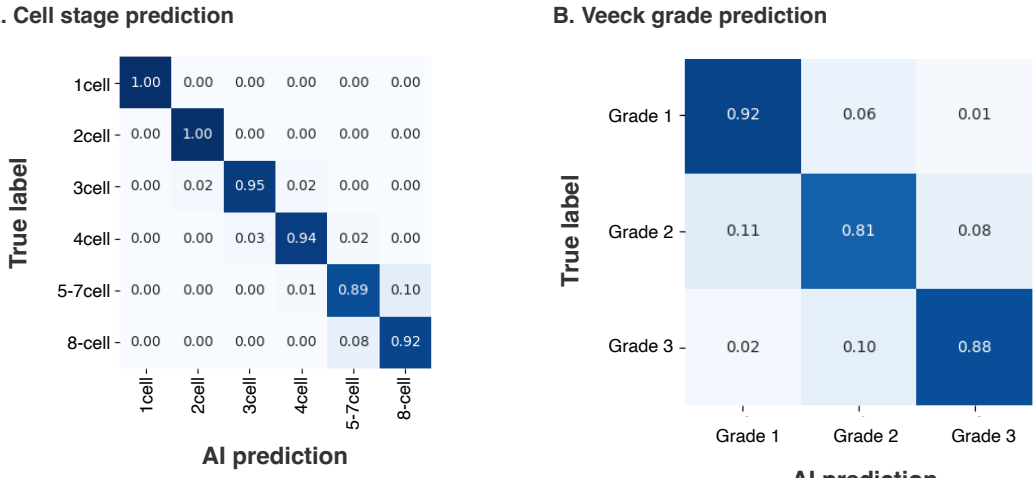


A. Absolute mean of SHAP values of the top 30 contributing features

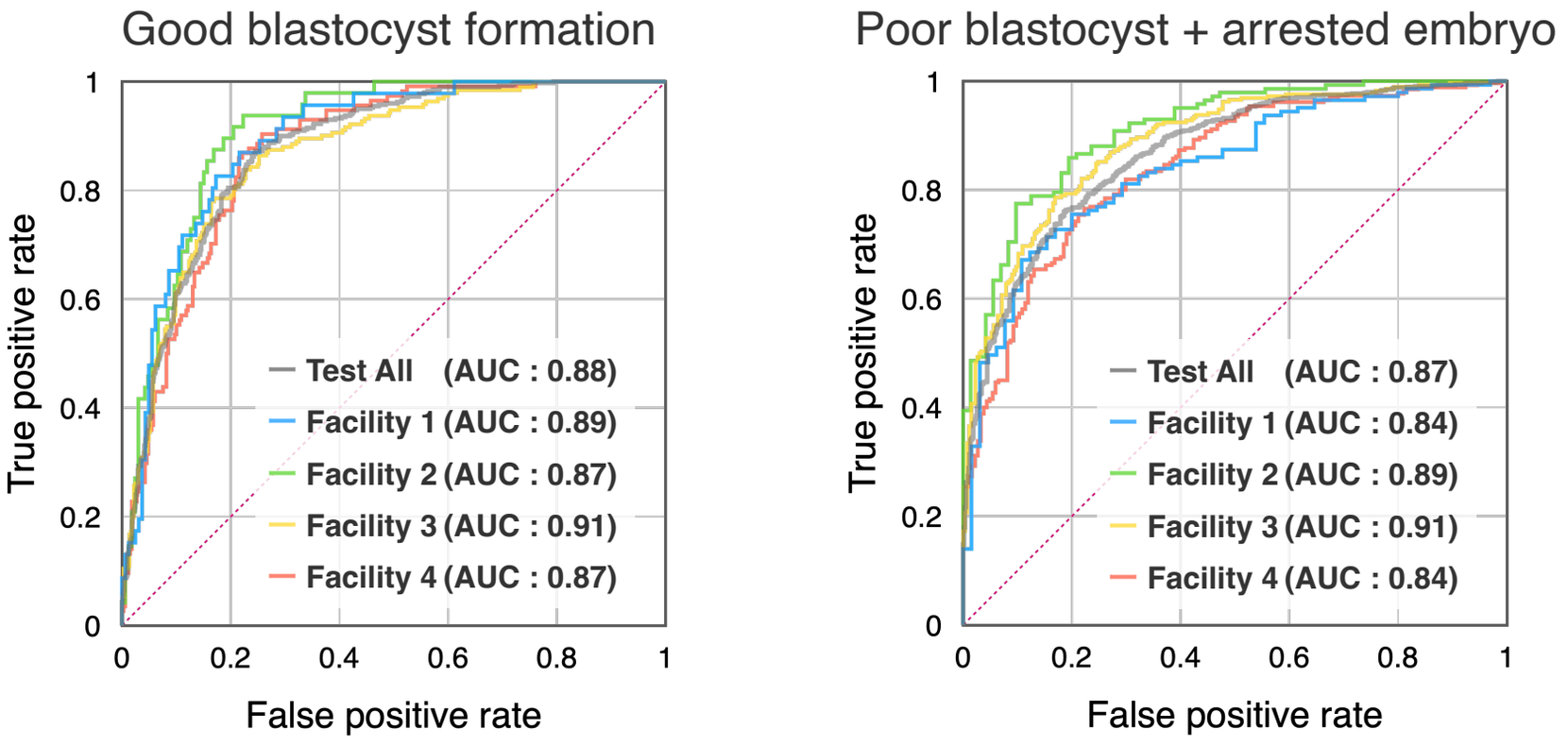


B. SHAP values and highly contributive features of the blastocyst formation prediction model

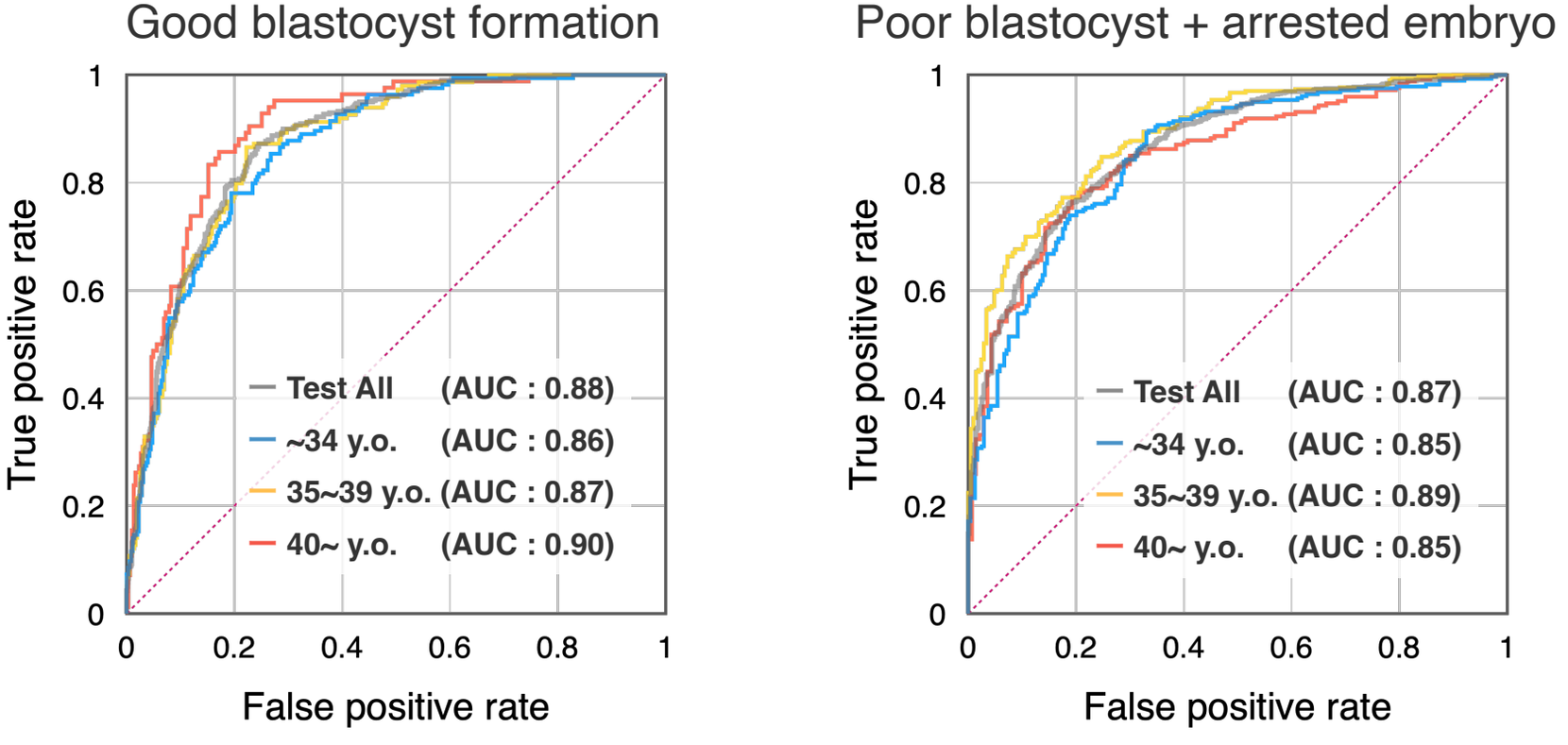



A. Cell stage prediction
True label
1cell
2cell
3cell
4cell
5-7cell
8-cell
1.00 0.00 0.00 0.00 0.00 0.00
0.00 1.00 0.00 0.00 0.00 0.00
0.00 0.02 0.95 0.02 0.00 0.00
0.00 0.00 0.03 0.94 0.02 0.00
0.00 0.00 0.00 0.01 0.89 0.10
0.00 0.00 0.00 0.00 0.08 0.92
1cell
2cell
3cell
4cell
5-7cell
8-cell
AI prediction
B. Veeck grade prediction
True label
Grade 1
Grade 2
Grade 3
0.92 0.06 0.01
0.11 0.81 0.08
0.02 0.10 0.88
Grade 1
Grade 2
Grade 3
AI prediction

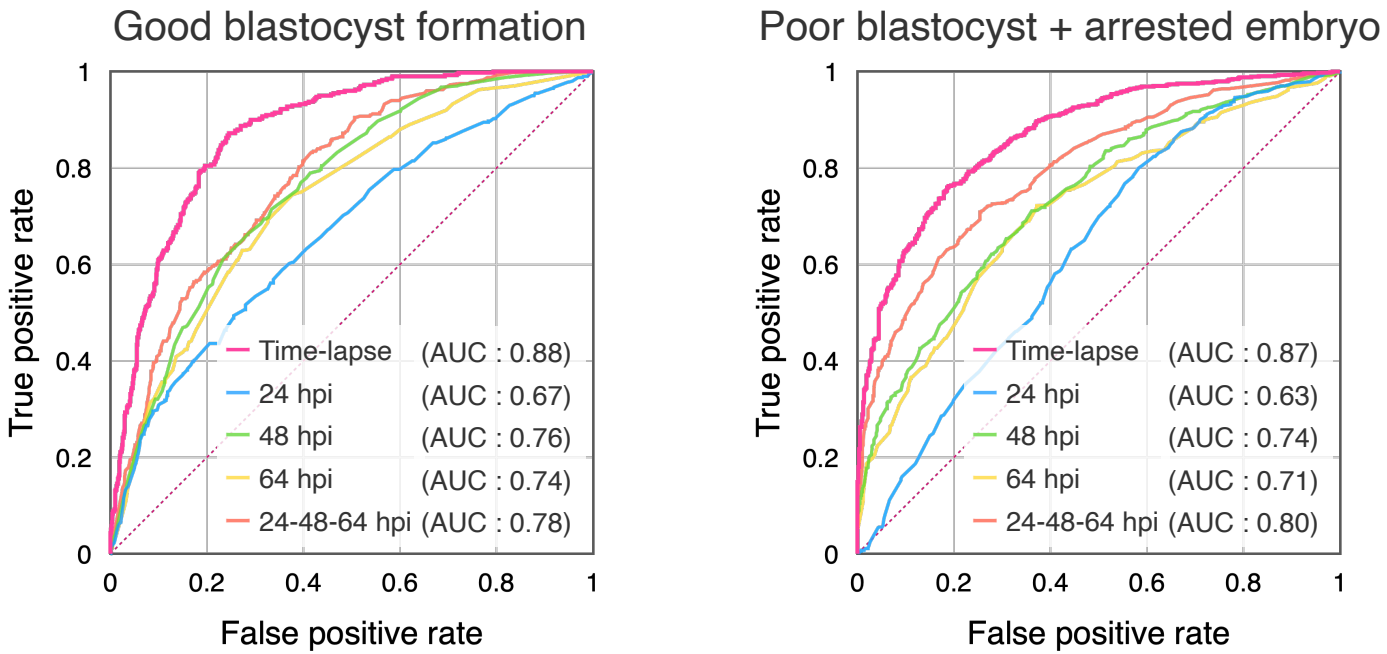
A. ROC subanalyses stratified by facility



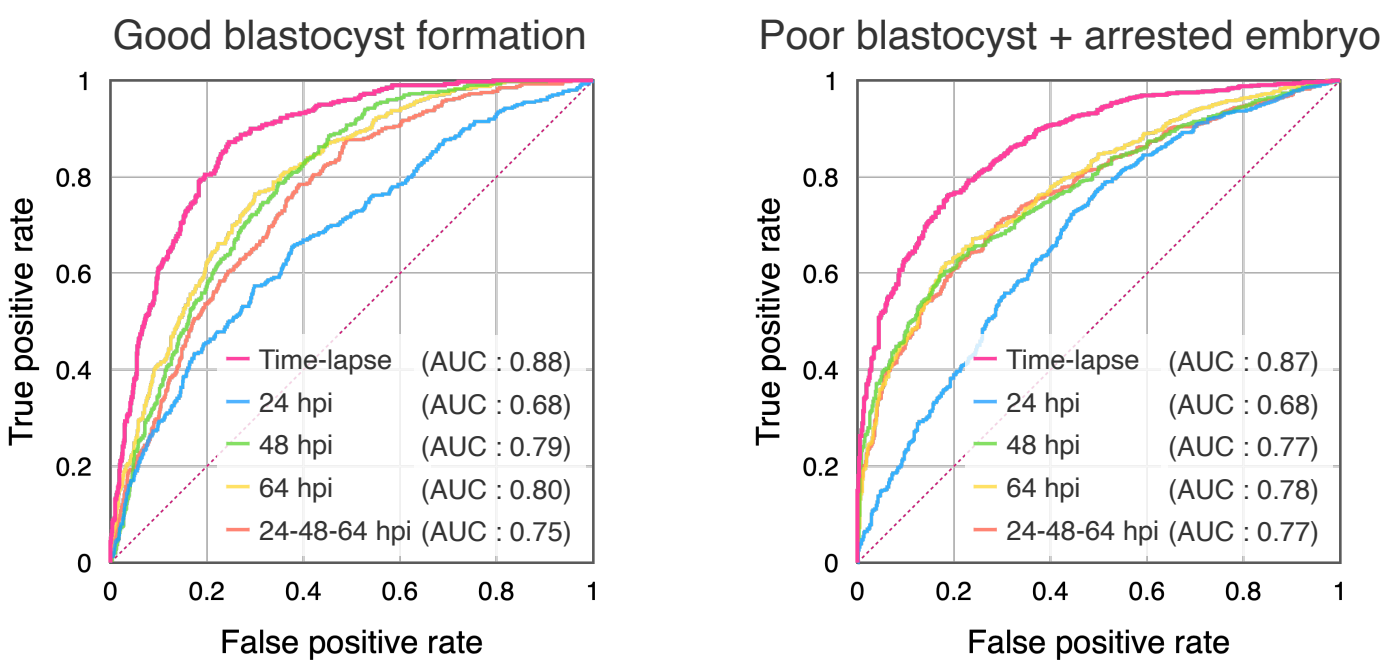
B. ROC subanalyses stratified by age group



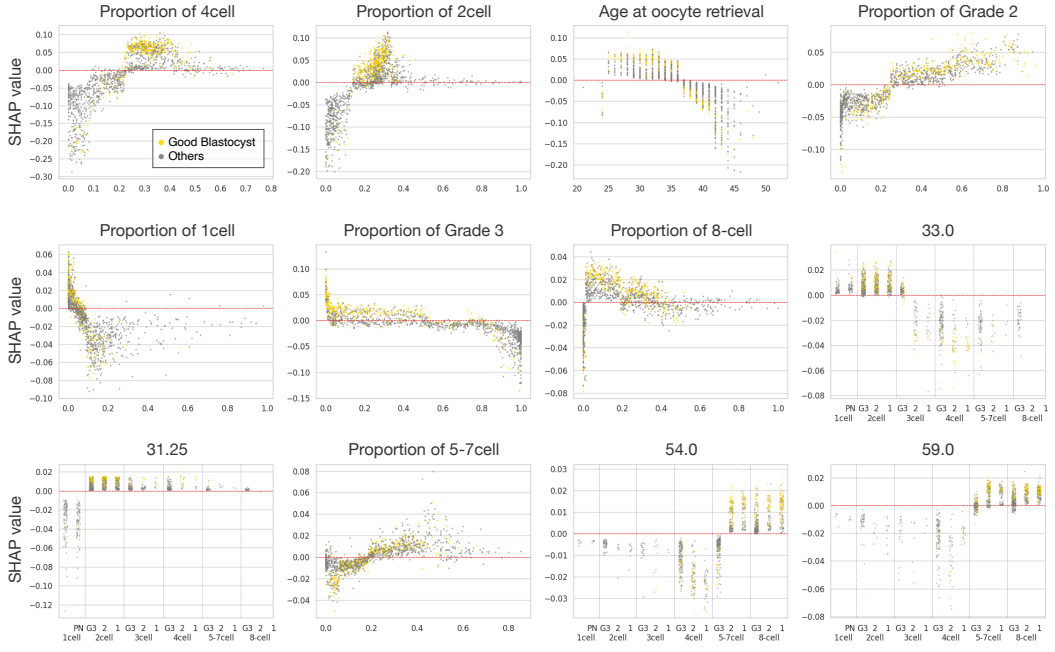
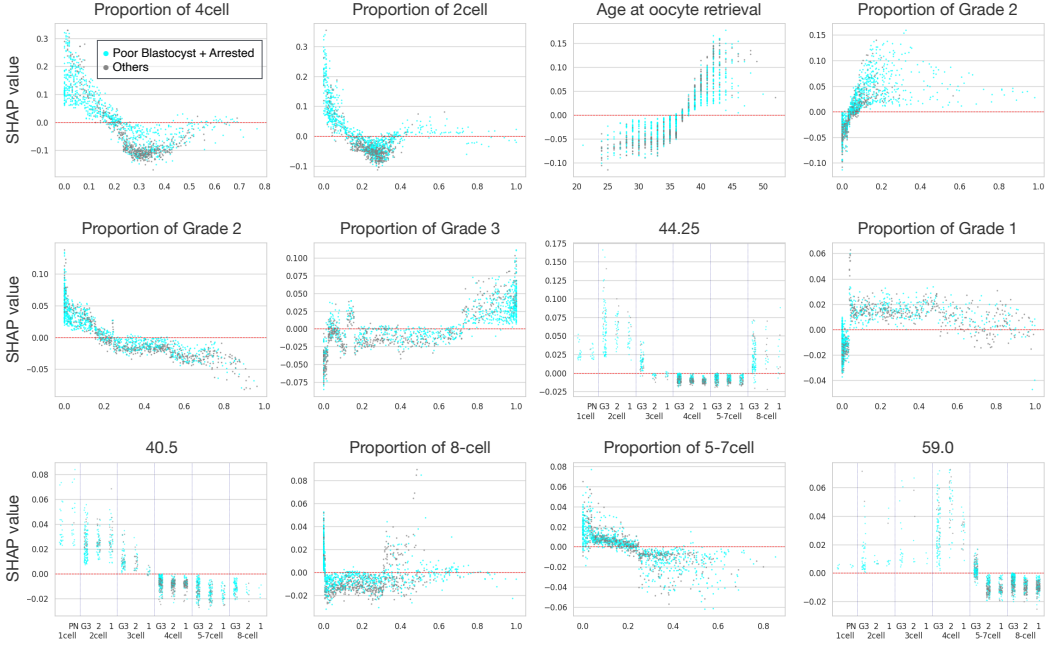
A. Comparison of time-lapse-based model with one or three label-based model



B. Comparison of time-lapse-based model with one of three image-based model



A. SHAP values and highly contributive features of the good blastocyst formation prediction model



B. SHAP values and highly contributive features of the poor blastocyst + arrested embryo prediction model