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

Study question: Can device-independent artificial intelligence (AI) models accurately predict blastocyst formation and quality by embryonic day 3 using time-lapse images and age?

Summary answer: The AI-based prediction models achieved receiver operating characteristic (ROC) area under the curve (AUC) ≥0.87 for predicting blastocyst formation, good blastocysts, and poor blastocyst + arrested embryos, demonstrating high accuracy at embryonic day 3.

What is known already: Previous blastocyst prediction models from embryonic day 3 time-lapse images and parameters have achieved moderate accuracy (ROC AUCs: 0.73–0.82). These models require labor-intensive manual parameter annotation or exhibit device dependence, limiting their generalizability and clinical usefulness.

Study design, size, duration: This was a retrospective, multicenter study using 7,111 eligible human embryos cultured in four types of time-lapse incubators across four facilities from May 2018 to December 2022. Additionally, 23,852 time-lapse and stereomicroscopic images of embryos during the same period were collected and annotated using 17-class supervised classification labels for the AI auto-annotator.

Participants/materials, setting, methods: Embryo culture time-lapse images and clinical data were retrospectively collected from Kyoto University Hospital and its cooperating facilities. This study included 2PN embryos cultured individually per well up to embryonic day 5 or 6. Degenerated embryos and those with poor Veeck grades (4 and 5) were excluded. Time-lapse images were recorded every 15 minutes between 24- and 64-hours post-insemination and preprocessed. Embryos with over 5% image loss or preprocessing errors were excluded. The final dataset was used to construct AI models for blastocyst formation and quality prediction. The dataset was split into 80% for training and cross-validation and 20% for independent testing. A pre-trained image classification AI (NASNet-A Large) was fine-tuned as AI auto-annotator, using 17-class supervised classification dataset including cell stages and Veeck grades. The AI auto-annotator converted time-lapse images into time-series labels. These labels, the proportions of each stage and grade, and the age at egg retrieval were analyzed using a gradient boosting algorithm (XGBoost) to predict blastocyst formation and quality. The performance of models was evaluated using ROC and precision-recall analyses with 5-fold validation and independent testing.

Main results and the role of chance: The AI auto-annotator achieved 85% accuracy in classifying embryo images into 17 classes, with 95% cell stage classification accuracy. The blastocyst formation prediction model had an ROC AUC of 0.86 for mean cross-validation and 0.87 for test. The good blastocyst formation prediction model had a ROC AUC of 0.88 for test. The poor blastocyst + arrested embryo prediction model had a ROC AUC of 0.87 for test and achieved an AUC of 0.90 for precision recall, meeting the target accuracy for clinical decision-making. The performance of models was evaluated through five-fold cross-

validation and independent testing. Subgroup analyses revealed minimal differences between facilities and age groups, demonstrating the generalizability of the model.

Limitations, reasons for caution: The predictive models used Gardner criteria, widely adopted but prone to inter-rater variability, potentially affecting consistency. Blastocyst evaluations at each facility were unmodified, and single-rater evaluations were avoided to prevent domain shift. The exclusion of embryos with defects or poor quality (Veeck grades 4 and 5) may limit the model's generalizability. Additionally, as this was a retrospective study, prospective validation is required to confirm the findings.

Wider implications of the findings: This versatile AI model enables more accurate embryo-by-embryo decisions by embryonic day 3, potentially guiding selection between continuing blastocyst culture versus early transfer/freezing, and thus improving individualized patient care in ART.

Study funding/competing interest(s): This work was supported by JSPS KAKENHI grant number JP23K08821. The authors have no conflicts of interest to declare.

Trial registration number: N/A.

Keywords

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

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

89

90

Introduction

Assisted reproductive technology (ART) plays an important role in infertility treatment. In 2021, 9% of all births in Japan resulted from ART (Katagiri *et al.*, 2024). 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 (Cameron *et al.*, 2020; Glujovsky *et al.*, 2022). According to the Japanese National Registry of ART ("the Japan Society of Obstetrics and Gynecology ART Databook 2021," n.d.; Katagiri *et al.*, 2024), 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 (Vlaisavljević *et al.*, 2001; Kovačič *et al.*, 2002; De Croo *et al.*, 2022) showed that extending culture to the blastocyst stage did not improve implantation or cumulative pregnancy rates in certain scenarios. Additionally, one report (Xiao *et al.*, 2019) 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 (Katagiri *et al.*, 2024).

In previous studies (Glujovsky *et al.*, 2022), 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

114 prediction of morphological quality on embryonic day 5 or 6 by embryonic day 3. Although
115 one report (Jiang *et al.*, 2022) showed that static morphological assessment on embryonic
116 day 3 could predict blastocyst formation, the receiver operating characteristic area under
117 the curve (ROC AUC) of 0.78 was considered insufficient for reliable clinical decision-making.

118 Since 2008, time-lapse embryo culture (Leung *et al.*, 2016) has enabled the evaluation of
119 morphokinetic parameters, leading to multiple studies reporting predictions of blastocyst
120 formation (Wong *et al.*, 2010; Cruz *et al.*, 2012; Dal Canto *et al.*, 2012; Chamayou *et al.*,
121 2013; Conaghan *et al.*, 2013; Athayde Wirka *et al.*, 2014; Milewski *et al.*, 2015; Yang *et al.*,
122 2015; Motato *et al.*, 2016; Mizobe *et al.*, 2018; Zaninovic *et al.*, 2019) and good blastocyst
123 formation (Kirkegaard *et al.*, 2013; Cetinkaya *et al.*, 2015; Storr *et al.*, 2015; Yang *et al.*, 2015;
124 Mizobe *et al.*, 2018) using these parameters. While day 4 predictions of blastocyst formation
125 achieved validated AUCs as high as 0.85 (Motato *et al.*, 2016), clinically useful predictions at
126 day 3 remained lower, with AUCs of 0.73–0.81 (Milewski *et al.*, 2015; Zaninovic *et al.*, 2019).

127 Recent advances have included artificial intelligence (AI) analysis of time-lapse images (Tran
128 *et al.*, 2019; Huang *et al.*, 2021; Lee *et al.*, 2021; Liao *et al.*, 2021; Barnes *et al.*, 2023), with
129 one reported AI-based model achieving an AUC of 0.82 for predicting blastocyst formation
130 on embryonic day 3 (Liao *et al.*, 2021). However, previously reported predictive models
131 (Milewski *et al.*, 2015; Zaninovic *et al.*, 2019; Liao *et al.*, 2021; Jiang *et al.*, 2022) face several
132 limitations. First, the prediction accuracy was insufficient for reliable clinical decisions
133 regarding whether extended blastocyst culture or early embryo transfer/freezing on
134 embryonic day 3. Next, the time-lapse parameters required laborious manual annotation
135 and exhibited interrater variability (Sundvall *et al.*, 2013). Last, AI prediction models often
136 lacked interpretability (“black box” issue) and were highly device-dependent, as they were

137 trained using images from only one type of time-lapse incubator (Tran *et al.*, 2019; Huang *et*
138 *al.*, 2021; Lee *et al.*, 2021; Liao *et al.*, 2021; Barnes *et al.*, 2023).

139 This study aimed to develop an accurate AI blastocyst prediction model on embryonic day 3
140 with high versatility and interpretability for clinical use. To achieve this objective, we
141 analyzed time-lapse images acquired from multiple time-lapse incubators at multiple
142 facilities, constructed models from human-interpretable features, and visualized the AI
143 decision basis using SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017;
144 Lundberg *et al.*, 2018).

145

146 Materials and Methods

147 Overview of the Blastocyst Prediction using AI

148 An overview is shown in Figure 1. We created a dataset for the blastocyst prediction model
149 (Figure 1A). We fine-tuned the ImageNet-1k pre-trained image analysis AI (NASNet-A Large
150 (Zoph *et al.*, 2018)), using a 17-class embryo image classification dataset to build an AI auto-
151 annotator (Figure 1B). AI extracted features from time-lapse images, and the age at oocyte
152 retrieval was analyzed using a gradient boosting method (XGBoost (Chen and Guestrin,
153 2016)) to build each prediction model for blastocyst formation, good blastocyst formation,
154 and poor blastocyst + arrested embryos (Figure 1E). Our AI blastocyst prediction model
155 predicted embryonic morphology on embryonic day 5 or 6 based on embryo culture time-
156 lapse images and age at oocyte retrieval.

157 Fine-Tuning the AI Auto-Annotator

158 A 17-class embryo image classification dataset (Figure 2A) was created from 23,852 embryo
159 images with 17-class supervised labels, including 1-cell stage with or without pronuclei (PN)
160 and 2-, 3-, 4-, 5–7-, and 8-cell stages with Veeck grades 1/2/3 (Veeck, 1988). The ImageNet-
161 1k pre-trained image AI, NASNet-A Large (Zoph *et al.*, 2018), was fine-tuned on this dataset
162 to create an AI auto-annotator (Supplementary Text 1). Time-lapse images were analyzed
163 using the AI auto-annotator and classified into 17-class morphological labels. The
164 proportions of each cell stage and Veeck grade were calculated in each embryo as features.
165 The Veeck grade was replaced by the mode frequency (or a better grade if the mode
166 frequency was the same) at each cell stage for each embryo.

167 **Data Collection and Preparation for the Blastocyst Prediction Dataset**

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

171 This study included 2PN embryos with cleavage that were time-lapse cultured, with at least
172 embryonic day 5, single embryo per well. Degenerated and poor embryos with Veeck grades
173 4 and 5 were excluded. Time-lapse images were collected every 15 min at 24–64 hours post-
174 insemination (hpi). Time-lapse images were collected up to 64 hpi because most day 3
175 embryo transfer/freezing were performed after 64 hpi. Time-lapse videos were
176 preprocessed and converted to JPG images of 331 × 331 pixels. Embryos with more than 5%
177 image loss were excluded due to missing data, debris, misplacement or significant
178 preprocessing errors. (Supplementary Text 1).

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

182 Overall, 8,043 fertilized embryos cultured in four types of time-lapse incubators in 2,792
183 oocyte retrieval cycles at four facilities were included. After exclusions, 7,111 embryos
184 remained. The dataset was randomly split into 80% for training and cross-validation and 20%
185 (1,422 embryos) for testing (Figure 1A).

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

190 **Building the AI Blastocyst Prediction Models**

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

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

202 To examine the superiority of time-lapse image analysis over static image analysis,
203 prediction models were created from a single image at 24, 48, or 64 hpi, or three images at
204 24, 48, and 64 hpi, assuming a once-a-day observation, and ROC AUC values were compared.

205 In the first type of model, the AI auto-annotator classified static images into morphological
206 annotation labels, and XGBoost built prediction models from single or three labels and age.
207 For the next type of model, the fine-tuned NASNet-A Large was trained on single images
208 (grayscale, 331 × 331 pixels) at 24, 48, and 64 hpi in the training and cross-validation dataset
209 to create prediction models. Additionally, because NASNet-A Large is a color image analysis
210 AI that allows up to three channels of input for each image, we created merged images using
211 24-, 48-, and 64-hpi images for each channel and trained NASNet-A Large on the merged
212 images to create the prediction models.

213

214 **SHAP-Based Model Interpretation**

215 In our prediction models, the AI auto-annotator classified embryo images into human-
216 recognizable morphological labels, and all features inputted to XGBoost were interpretable.
217 However, because XGBoost has dozens to hundreds of decision trees, it is difficult to directly
218 interpret the basis for the prediction models. Therefore, we used SHAP (Lundberg and Lee,
219 2017; Lundberg *et al.*, 2018) to analyze the contribution and predictive direction of each
220 feature in the test dataset. Because SHAP values are positive and negative for positive and
221 negative predictions, respectively, we calculated the absolute mean of SHAP values for each

222 feature and extracted the highly contributing features. The correlation between the value
223 and SHAP value of each contributing feature was visualized and used to interpret the models.

224

225 **Statistical Analyses**

226 The model performance was evaluated using ROC and PR analyses and AUCs. The analyses
227 were performed using numpy 1.21.6, scipy 1.7.1, and sklearn 1.0.2 on Python 3.7.15. In the
228 comparison of blastocyst prediction models from time-lapse images and those from single or
229 few embryo images, ROC AUCs were compared, and p-values were calculated using the
230 DeLong test. The DeLong test was performed using the fast DeLong method in Python
231 (https://github.com/yandexdataschool/roc_comparison) (DeLong *et al.*, 1988; Sun and Xu,
232 2014). The significance level was set at $p<0.05$; however, because of nine-fold tests, the
233 significance level was set at $p<0.0055$ with Bonferroni correction.

234

235 **Ethical Considerations**

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

240

241 **Results**

242 **AI Auto-Annotator Performance**

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

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

252 Major time-lapse parameters were calculated from the AI annotation labels and compared
253 with those in previous reports (Cruz *et al.*, 2012; Dal Canto *et al.*, 2012; Milewski *et al.*, 2015;
254 Motato *et al.*, 2016; Canosa *et al.*, 2020) (Supplementary Table 1). AI-calculated t2 (time to
255 the 2-cell stage) was significantly smaller in the blastocyst formation group than in the
256 arrested embryo group (median: 24.8 vs. 27.0 hpi, $p<0.001$) and was within the mean or
257 median values in previous reports.

258 **Blastocyst Formation Prediction Model**

259 Using XGBoost and recursive feature elimination, 88 features were selected for the
260 blastocyst formation prediction model (Supplementary Table 2). The model achieved a test
261 AUC of 0.87 and a five-fold cross-validation mean AUC of 0.86 (Figure 3A). With a Youden's
262 index (Youden, 1950) threshold of 0.49, the test accuracy reached 0.79. These results
263 demonstrated the model's robustness in cross-validation and testing.

264 **Good Blastocyst Formation Prediction Model**

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

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

269 The poor blastocyst + arrested embryo prediction model (Figure 3C), using 76 features,
270 achieved a test AUC of 0.87. Its PR AUC reached 0.90, indicating superior positive predictive
271 accuracy compared to other models and meeting the target performance criteria. Designed
272 as a clinical tool, this model assists in selecting early embryo transfer or freezing on
273 embryonic day 3 instead of blastocyst culture. Adjusting the decision threshold further
274 personalizes the clinical application, enabling a focus on achieving a 90% positive predictive
275 value (PPV) at 60% sensitivity or an 80% PPV at 80% sensitivity (Figure 3D).

276 **Subgroup Analyses**

277 Subgroup analyses by facility for all models indicated no large differences between facilities,
278 and the blastocyst formation prediction model was able to predict with an accuracy of AUC
279 ≥0.83 for each facility (Figure 4A and Supplementary Figure 2A). Including age at oocyte
280 retrieval as a feature, subgroup analyses by age group showed small differences between
281 groups, and predictions were possible irrespective of the age group (Figure 4B and
282 Supplementary Figure 2B). These findings highlight the robustness of our models.

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

284 The AUC of the blastocyst formation prediction model built using XGBoost based on AI auto-
285 annotated labels from single or three images and age was 0.63–0.78; however, the AUC of
286 the model based on time-lapse image analysis was significantly higher, at 0.87 ($p<1.0\times10^{-5}$

287 for all comparisons; Figure 5A). The blastocyst formation prediction models based on image-
288 only analysis by NASNet-A Large had an AUC of 0.64–0.76, and the prediction model based
289 on time-lapse image analysis had a significantly higher AUC, at 0.87 ($p<1.0\times10^{-5}$ for all
290 comparisons) (Figure 5B). The models based on time-lapse image analysis were also
291 significantly better ($p<1.0\times10^{-5}$ for all comparisons) for good blastocyst formation prediction
292 and poor blastocyst + arrested embryo prediction (Supplementary Figure 3).

293 **Feature Contribution and SHAP Analysis**

294 The top 30 features for each model, ranked by absolute mean SHAP values (Figure 6A),
295 consistently included proportions (2-cell stage, 4-cell stage, Veeck grade 2, etc.) and age at
296 oocyte retrieval. While the time points of high contribution varied across models, a wide
297 range of features within the 24–64 hpi broadly contributed to the models.

298 The correlation between annotation labels and SHAP values (Figure 6B) at each time point in
299 the blastocyst formation prediction model indicated that whether each feature contributes
300 to a positive or negative prediction often varies at a single point because XGBoost deals with
301 the 17-class labels as discrete-value variables rather than categorical variables. The 5–7-cell
302 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
303 stage or less at 41.75 hpi were found to contribute to positive predictions. Focusing on 41.75
304 hpi, the 5-cell or more stage due to overestimation of poor-grade embryos contributed to
305 negative predictions, whereas the 3-cell or less stage, which was the most frequently
306 arrested stage, incorrectly contributed to positive predictions. In contrast, at 42.75 hpi, the
307 3-cell or less stage contributed appropriately to negative predictions, and the entire model
308 worked well.

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

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

315

316 ## Discussion

317 ### Main Findings

318 In this study, 24–64-hpi time-lapse images recorded at various incubators and ages at oocyte
319 retrieval from multiple centers were analyzed using AI to build models for predicting
320 blastocyst morphology after the embryonic day 5. The AUCs for the blastocyst formation,
321 good blastocyst formation, and poor blastocyst + arrested embryo prediction models were
322 0.87, 0.88, and 0.87, respectively. The positive predictive accuracy of the poor blastocyst +
323 arrested embryo prediction model, which is envisioned as a clinical tool for deciding whether
324 blastocyst culture should be continued, or early embryo freezing/transfer should be
325 performed, achieved a PR AUC of 0.90, meeting the target accuracy.

326 ### Performance and Versatility of the AI Auto-Annotator

327 The newly developed AI auto-annotator offers several advantages over previously published
328 models. In particular, it is compatible with various types of incubators, attains the highest
329 level of cell stage classification accuracy without labor-intensive manual annotation, and

330 incorporates additional qualitative evaluations, including pronuclearity and Veeck grade. Our
331 cell stage accuracy of 95% compares favorably with previous studies (Khan *et al.*, 2016;
332 Dirvanauskas *et al.*, 2019; Raudonis *et al.*, 2019; Leahy *et al.*, 2020; Danardono *et al.*, 2022).
333 This performance was particularly remarkable considering that we included fragmented
334 embryos (Veeck grade 3) and more difficult-to-classify stages such as 3- and 5-7-cell. A
335 previous study reported a slightly higher accuracy (Dirvanauskas *et al.*, 2019), but their
336 model included fewer classification categories and a limited set of embryos, which may not
337 reflect real-world diversity. In contrast, our approach includes a broader range of embryo
338 states, including those with fragmentation, making it more generalizable to clinical practice.

339 **Accuracy and Generalizability of the Prediction Models**

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

343 Whereas prior day 3 blastocyst formation prediction models achieved validated ROC AUCs
344 ranging from 0.73 to 0.82 (Milewski *et al.*, 2015; Zaninovic *et al.*, 2019; Liao *et al.*, 2021; Jiang
345 *et al.*, 2022), our model demonstrated superior performance with an AUC of 0.87. Similarly,
346 our good blastocyst formation model achieved an ROC AUC of 0.88, improving on reported
347 unvalidated AUCs of 0.73–0.82 (Kirkegaard *et al.*, 2013; Cetinkaya *et al.*, 2015; Storr *et al.*,
348 2015) . Moreover, we successfully created a poor blastocyst + arrested embryo model with a
349 PR AUC of 0.90, indicating its utility for identifying embryos unlikely to develop further. This
350 prediction can guide clinicians in deciding whether to proceed with day 3 transfer or freezing
351 the embryo rather than risk further culture.

352 Most previous models relied on time-lapse parameters or images captured within 24 hpi
353 (Wong *et al.*, 2010; Cruz *et al.*, 2012; Dal Canto *et al.*, 2012; Chamayou *et al.*, 2013;
354 Conaghan *et al.*, 2013; Kirkegaard *et al.*, 2013; Athayde Wirka *et al.*, 2014; Cetinkaya *et al.*,
355 2015; Storr *et al.*, 2015; Yang *et al.*, 2015; Milewski *et al.*, 2015; Mizobe *et al.*, 2018;
356 Zaninovic *et al.*, 2019; Liao *et al.*, 2021). However, our study included only images of the
357 embryos from 24 to 64 hpi, which are routinely recorded at most facilities, ensuring broader
358 applicability while maintaining superior accuracy.

359 In contrast to previous AI models, which were device-dependent (Tran *et al.*, 2019; Huang *et*
360 *al.*, 2021; Lee *et al.*, 2021; Liao *et al.*, 2021), our study analyzed images from four types of
361 incubators across four facilities. By using a highly adaptable pre-processing method and a
362 versatile AI auto-annotator, our models effectively managed variations such as aspect ratios,
363 embryo area percentages, imaging contrasts, resolutions, and well geometries.

364 Data were collected irrespective of oocyte retrieval age, insemination method, or culture
365 medium, creating heterogeneity across cases. Nonetheless, subgroup analyses showed
366 minimal differences in AUC across facilities and age groups, confirming the model's
367 generalizability. These findings suggest that our models are versatile, robust, and suitable for
368 diverse clinical settings.

369

370 **Interpretability of the prediction models**

371 XGBoost outperformed Long Short-Term Memory models (data not shown) by analyzing
372 features in parallel, effectively capturing temporal aspects of morphokinetic development
373 from 24–64 hpi. SHAP analysis revealed how features, including human-recognizable

374 parameters and embryo age, contributed to predictions. Importantly, localized
375 misclassifications, such as overestimated cell stages in fragmented embryos, did not
376 compromise overall model accuracy due to integration with more reliable features.

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

382 **Limitations**

383 Some limitations should be considered. First, we relied on the Gardner criteria (Gardner,
384 1999; Gardner *et al.*, 2000) for blastocyst grading. While widely used, these criteria are
385 subjective and prone to inter-rater variability (Storr *et al.*, 2017). To minimize domain shift,
386 grading was conducted independently at each facility without standardization.

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

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

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

395

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

437 Athayde Wirka K, Chen AA, Conaghan J, Ivani K, Gvakharia M, Behr B, Suraj V, Tan L, Shen S.
438 Atypical embryo phenotypes identified by time-lapse microscopy: High prevalence and
439 association with embryo development. *Fertil Steril* 2014;**101**:1637–1648.e5.

440 Barnes J, Brendel M, Gao VR, Rajendran S, Kim J, Li Q, Malmsten JE, Sierra JT, Zisimopoulos P,
441 Sigaras A, *et al.* A non-invasive artificial intelligence approach for the prediction of
442 human blastocyst ploidy: a retrospective model development and validation study.
443 *Lancet Digit Heal* 2023;**5**:e28–e40.

444 Cameron NJ, Bhattacharya S, McLernon DJ. Cumulative live birth rates following blastocyst-
445 versus cleavage-stage embryo transfer in the first complete cycle of IVF: A population-
446 based retrospective cohort study. *Hum Reprod* 2020;**35**:2365–2374.

447 Canosa S, Bergandi L, Macrì C, Charrier L, Paschero C, Carosso A, Segni N Di, Silvagno F,
448 Gennarelli G, Benedetto C, *et al.* Morphokinetic analysis of cleavage stage embryos and
449 assessment of specific gene expression in cumulus cells independently predict human
450 embryo development to expanded blastocyst: a preliminary study. *J Assist Reprod*
451 *Genet* 2020;**37**:1409–1420.

452 Cetinkaya M, Pirkevi C, Yelke H, Colakoglu YK, Atayurt Z, Kahraman S. Relative kinetic
453 expressions defining cleavage synchronicity are better predictors of blastocyst
454 formation and quality than absolute time points. *J Assist Reprod Genet* 2015;**32**:27–35.

455 Chamayou S, Patrizio P, Storaci G, Tomaselli V, Alecci C, Ragolia C, Crescenzo C, Guglielmino
456 A. The use of morphokinetic parameters to select all embryos with full capacity to
457 implant. *J Assist Reprod Genet* 2013;**30**:703–710.

458 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proc ACM SIGKDD Int Conf*
459 *Knowl Discov Data Min* 2016;**13**-**17**-**Augu**:785–794.

460 Conaghan J, Chen AA, Willman SP, Ivani K, Chenette PE, Boostanfar R, Baker VL, Adamson GD,
461 Abusief ME, Gvakharia M, *et al.* Improving embryo selection using a computer-
462 automated time-lapse image analysis test plus day 3 morphology: results from a
463 prospective multicenter trial. *Fertil Steril* 2013;**100**:412–419.e5.

464 Croo I De, Colman R, Sutter P De, Stoop D, Tilleman K. No difference in cumulative live birth
465 rates between cleavage versus blastocyst transfer in patients with four or fewer
466 zygotes: results from a retrospective study. *Hum Reprod Open* 2022;**2022**:1–9.

467 Cruz M, Garrido N, Herrero J, Pérez-Cano I, Muñoz M, Meseguer M. Timing of cell division in
468 human cleavage-stage embryos is linked with blastocyst formation and quality. *Reprod*
469 *Biomed Online* 2012;**25**:371–381.

470 Dal Canto M, Coticchio G, Mignini Renzini M, Ponti E De, Novara PV, Brambillasca F, Comi R,
471 Fadini R. Cleavage kinetics analysis of human embryos predicts development to
472 blastocyst and implantation. *Reprod Biomed Online* 2012;**25**:474–480

473 Danardono GB, Erwin A, Purnama J, Handayani N, Polim AA, Boediono A, Sini I. A
474 Homogeneous Ensemble of Robust Pre-defined Neural Network Enables Automated
475 Annotation of Human Embryo Morphokinetics. *J Reprod Infertil* 2022;**23**:250–256.

476 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more
477 correlated receiver operating characteristic curves: a nonparametric approach.
478 *Biometrics* 1988;**44**:837–845.

479 Dirvanauskas D, Maskeliunas R, Raudonis V, Damasevicius R. Embryo development stage
480 prediction algorithm for automated time lapse incubators. *Comput Methods Programs*
481 *Biomed* 2019;**177**:161–174.

482 Gardner DK. In-vitro culture of human blastocyst. *Towar Reprod Certain Infertil Genet*
483 *beyond 1999* 1999;, p. 378–388. Parthenon Press.

484 Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects
485 implantation and pregnancy outcome: Towards a single blastocyst transfer. *Fertil Steril*
486 2000;**73**:1155–1158.

487 Glujovsky D, Quinteiro Retamar AM, Alvarez Sedo CR, Ciapponi A, Cornelisse S, Blake D.
488 Cleavage-stage versus blastocyst-stage embryo transfer in assisted reproductive
489 technology. *Cochrane database Syst Rev* 2022;**5**:CD002118.

490 Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can
491 predict embryo ploidy status based on time-lapse data. *Reprod Biol Endocrinol*
492 2021;**19**:1–10.

493 Jiang X, Cai J, Liu L, Liu Z, Wang W, Chen J, Yang C, Geng J, Ma C, Ren J. Does conventional
494 morphological evaluation still play a role in predicting blastocyst formation? *Reprod Biol*
495 *Endocrinol* 2022;**20**:1–12.

496 Katagiri Y, Jwa SC, Kuwahara A, Iwasa T, On M, Kato K, Kishi H, Kuwabara Y, Taniguchi F,
497 Harada M, *et al.* Assisted reproductive technology in Japan: A summary report for 2021
498 by the Ethics Committee of the Japan Society of Obstetrics and Gynecology. *Reprod*
499 *Med Biol* 2024;**23**:e12552.

500 Khan A, Gould S, Salzmann M. Deep convolutional neural networks for human embryonic cell
501 counting. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes*
502 *Bioinformatics)* 2016;**9913 LNCS**:339–348.

503 Kirkegaard K, Kesmodel US, Hindkjær JJ, Ingerslev HJ. Time-lapse parameters as predictors of
504 blastocyst development and pregnancy outcome in embryos from good prognosis
505 patients: A prospective cohort study. *Hum Reprod* 2013;**28**:2643–2651.

506 Kovačič B, Vlaisavljevi V, Reljič M, Gavri Lovrec V. Clinical outcome of day 2 versus day 5
507 transfer in cycles with one or two developed embryos. *Fertil Steril* 2002;**77**:529–536.

508 Leahy BD, Jang W-D, Yang HY, Struyven R, Wei D, Sun Z, Lee KR, Royston C, Cam L, Kalma Y,
509 *et al.* Automated Measurements of Key Morphological Features of Human Embryos for
510 IVF. *Med Image Comput Comput Assist Interv* 2020;**12265**:25–35.

511 Lee CI, Su YR, Chen CH, Chang TA, Kuo EES, Zheng WL, Hsieh WT, Huang CC, Lee MS, Liu M.
512 End-to-end deep learning for recognition of ploidy status using time-lapse videos. *J*
513 *Assist Reprod Genet* 2021;**38**:1655–1663.

514 Leung AS, Son WY, Dahan MH. Time-lapse imaging of embryos: current evidence supporting
515 its use. *Expert Rev Med Devices* 2016;**13**:881–883.

516 Liao Q, Zhang Q, Feng X, Huang H, Xu H, Tian B, Liu J, Yu Q, Guo N, Liu Q, *et al.* Development
517 of deep learning algorithms for predicting blastocyst formation and quality by time-
518 lapse monitoring. *Commun Biol* 2021;**4**:415.

519 Lundberg SM, Erion GG, Lee S-I. Consistent Individualized Feature Attribution for Tree
520 Ensembles. 2018;Available from: http://arxiv.org/abs/1802.03888.

521 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf*
522 *Process Syst* 2017;4766–4775.

523 Milewski R, Kuć P, Kuczyńska A, Stankiewicz B, Łukaszuk K, Kuczyński W. A predictive model
524 for blastocyst formation based on morphokinetic parameters in time-lapse monitoring
525 of embryo development. *J Assist Reprod Genet* 2015;**32**:571–579.

526 Mizobe Y, Tokunaga M, Oya N, Iwakiri R, Yoshida N, Sato Y, Onoue N, Ezono Y. Synchrony of
527 the first division as an index of the blastocyst formation rate during embryonic
528 development. *Reprod Med Biol* 2018;**17**:64–70.

529 Motato Y, los Santos MJ de, Escriba MJ, Ruiz BA, Remohí J, Meseguer M. Morphokinetic
530 analysis and embryonic prediction for blastocyst formation through an integrated time-
531 lapse system. *Fertil Steril* 2016;**105**:376–384.e9.

532 Raudonis V, Paulauskaite-Taraseviciene A, Sutiene K, Jonaitis D. Towards the automation of
533 early-stage human embryo development detection. *Biomed Eng Online* 2019;**18**:1–20.
534 BioMed Central.

535 Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer
536 agreement between embryologists during selection of a single Day 5 embryo for
537 transfer: A multicenter study. *Hum Reprod* 2017;**32**:307–314.

538 Storr A, Venetis CA, Cooke S, Susetio D, Kilani S, Ledger W. Morphokinetic parameters using
539 time-lapse technology and day 5 embryo quality: a prospective cohort study. *J Assist*
540 *Reprod Genet* 2015;**32**:1151–1160.

541 Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under
542 Correlated Receiver Operating Characteristic Curves. *IEEE Signal Process Lett* [Internet]
543 2014;**21**:1389–1393.

544 Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability
545 of time-lapse annotations. *Hum Reprod* 2013;**28**:3215–3221.

546 the Japan Society of Obstetrics and Gynecology ART Databook 2021Available from:
547 https://www.jsog.or.jp/activity/art/2021_JSOG-ART.pdf.

548 Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart
549 pregnancy following time-lapse incubation and blastocyst transfer. *Hum Reprod*
550 2019;**34**:1011–1018.

551 Veeck LL. Oocyte Assessment and Biological Performance. *Ann N Y Acad Sci* 1988;**541**:259–
552 274.

553 Vlaisavljević V, Kovačič B, Reljič M, Lovrec VG, Sajko MČ. Is there any benefit from the
554 culture of a single oocyte to a blastocyst-stage embryo in unstimulated cycles? *Hum*
555 *Reprod* 2001;**16**:2379–2383.

556 Wong CC, Loewke KE, Bossert NL, Behr B, Jonge CJ De, Baer TM, Pera RAR. Non-invasive
557 imaging of human embryos before embryonic genome activation predicts development
558 to the blastocyst stage. *Nat Biotechnol* 2010;**28**:1115–1121. Nature Publishing Group.

559 Xiao JS, Healey M, Talmor A, Vollenhoven B. When only one embryo is available, is it better
560 to transfer on Day 3 or to grow on? *Reprod Biomed Online* 2019;**39**:916–923.

561 Yang ST, Shi JX, Gong F, Zhang SP, Lu CF, Tan K, Leng LZ, Hao M, He H, Gu YF, *et al.* Cleavage
562 pattern predicts developmental potential of day 3 human embryos produced by IVF.
563 *Reprod Biomed Online* 2015;**30**:625–634.

564 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;**3**:32–35.

565 Zaninovic N, Nohales M, Zhan Q, los Santos ZMJ de, Sierra J, Rosenwaks Z, Meseguer M. A
566 comparison of morphokinetic markers predicting blastocyst formation and implantation
567 potential from two large clinical data sets. *J Assist Reprod Genet* 2019;**36**:637–646.

568 Zoph B, Vasudevan V, Shlens J, Le Q V. Learning Transferable Architectures for Scalable
569 Image Recognition. *2018 IEEE/CVF Conf Comput Vis Pattern Recognit* [Internet] 2018;, p.
570 8697–8710. IEEEAvailable from: https://ieeexplore.ieee.org/document/8579005/.

571

572 Athayde Wirka K, Chen AA, Conaghan J, Ivani K, Gvakharia M, Behr B, Suraj V, Tan L, Shen S.
573 Atypical embryo phenotypes identified by time-lapse microscopy: High prevalence and
574 association with embryo development. *Fertil Steril* 2014;**101**:1637-1648.e5. Elsevier Inc.

575 Barnes J, Brendel M, Gao VR, Rajendran S, Kim J, Li Q, Malmsten JE, Sierra JT, Zisimopoulos P,
576 Sigaras A, *et al.* A non-invasive artificial intelligence approach for the prediction of
577 human blastocyst ploidy: a retrospective model development and validation study.
578 *Lancet Digit Heal* 2023;**5**:e28–e40.

579 Cameron NJ, Bhattacharya S, McLernon DJ. Cumulative live birth rates following blastocyst-
580 versus cleavage-stage embryo transfer in the first complete cycle of IVF: A population-
581 based retrospective cohort study. *Hum Reprod* 2020;**35**:2365–2374.

582 Canosa S, Bergandi L, Macrì C, Charrier L, Paschero C, Carosso A, Segni N Di, Silvagno F,
583 Gennarelli G, Benedetto C, *et al.* Morphokinetic analysis of cleavage stage embryos and
584 assessment of specific gene expression in cumulus cells independently predict human
585 embryo development to expanded blastocyst: a preliminary study. *J Assist Reprod*
586 *Genet* 2020;**37**:1409–1420.

587 Cetinkaya M, Pirkevi C, Yelke H, Colakoglu YK, Atayurt Z, Kahraman S. Relative kinetic
588 expressions defining cleavage synchronicity are better predictors of blastocyst
589 formation and quality than absolute time points. *J Assist Reprod Genet* 2015;**32**:27–35.

590 Chamayou S, Patrizio P, Storaci G, Tomaselli V, Alecci C, Ragolia C, Crescenzo C, Guglielmino
591 A. The use of morphokinetic parameters to select all embryos with full capacity to
592 implant. *J Assist Reprod Genet* 2013;**30**:703–710.

593 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proc ACM SIGKDD Int Conf*
594 *Knowl Discov Data Min* 2016;**13**-**17**-**Augu**:785–794.

595 Conaghan J, Chen AA, Willman SP, Ivani K, Chenette PE, Boostanfar R, Baker VL, Adamson GD,
596 Abusief ME, Gvakharia M, *et al.* Improving embryo selection using a computer-
597 automated time-lapse image analysis test plus day 3 morphology: results from a
598 prospective multicenter trial. *Fertil Steril* 2013;**100**:412-9.e5.

599 Croo I De, Colman R, Sutter P De, Stoop D, Tilleman K. No difference in cumulative live birth
600 rates between cleavage versus blastocyst transfer in patients with four or fewer
601 zygotes: results from a retrospective study. *Hum Reprod Open* 2022;**2022**:1–9.

602 Cruz M, Garrido N, Herrero J, Pérez-Cano I, Muñoz M, Meseguer M. Timing of cell division in
603 human cleavage-stage embryos is linked with blastocyst formation and quality. *Reprod*
604 *Biomed Online* 2012;**25**:371–381.

605 Dal Canto M, Coticchio G, Mignini Renzini M, Ponti E De, Novara PV, Brambillasca F, Comi R,
606 Fadini R. Cleavage kinetics analysis of human embryos predicts development to
607 blastocyst and implantation. *Reprod Biomed Online* [Internet] 2012;**25**:474–480.
608 Reproductive Healthcare Ltd.

609 Danardono GB, Erwin A, Purnama J, Handayani N, Polim AA, Boediono A, Sini I. A
610 Homogeneous Ensemble of Robust Pre-defined Neural Network Enables Automated
611 Annotation of Human Embryo Morphokinetics. *J Reprod Infertil* 2022;**23**:250–256.

612 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more
613 correlated receiver operating characteristic curves: a nonparametric approach.
614 *Biometrics* 1988;**44**:837–845.

615 Dirvanauskas D, Maskeliunas R, Raudonis V, Damasevicius R. Embryo development stage
616 prediction algorithm for automated time lapse incubators. *Comput Methods Programs*
617 *Biomed* 2019;**177**:161–174. Elsevier B.V.

618 Gardner DK. In-vitro culture of human blastocyst. *Towar Reprod Certain Infertil Genet*
619 *beyond 1999* 1999;, p. 378–388. Parthenon Press.

620 Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects
621 implantation and pregnancy outcome: Towards a single blastocyst transfer. *Fertil Steril*
622 2000;**73**:1155–1158.

623 Glujovsky D, Quinteiro Retamar AM, Alvarez Sedo CR, Ciapponi A, Cornelisse S, Blake D.
624 Cleavage-stage versus blastocyst-stage embryo transfer in assisted reproductive
625 technology. *Cochrane database Syst Rev* 2022;**5**:CD002118.

626 Huang B, Tan W, Li Z, Jin L. An artificial intelligence model (euploid prediction algorithm) can
627 predict embryo ploidy status based on time-lapse data. *Reprod Biol Endocrinol*
628 2021;**19**:1–10. BioMed Central.

629 Jiang X, Cai J, Liu L, Liu Z, Wang W, Chen J, Yang C, Geng J, Ma C, Ren J. Does conventional
630 morphological evaluation still play a role in predicting blastocyst formation? *Reprod Biol*
631 *Endocrinol* 2022;**20**:1–12. BioMed Central.

632 Katagiri Y, Jwa SC, Kuwahara A, Iwasa T, On M, Kato K, Kishi H, Kuwabara Y, Taniguchi F,
633 Harada M, *et al.* Assisted reproductive technology in Japan: A summary report for 2021
634 by the Ethics Committee of the Japan Society of Obstetrics and Gynecology. *Reprod*
635 *Med Biol* 2024;**23**:e12552.

636 Khan A, Gould S, Salzmann M. Deep convolutional neural networks for human embryonic cell
637 counting. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes*
638 *Bioinformatics)* 2016;**9913 LNCS**:339–348.

639 Kirkegaard K, Kesmodel US, Hindkjær JJ, Ingerslev HJ. Time-lapse parameters as predictors of
640 blastocyst development and pregnancy outcome in embryos from good prognosis
641 patients: A prospective cohort study. *Hum Reprod* 2013;**28**:2643–2651.

642 Kovačič B, Vlaisavljevi V, Reljič M, Gavri Lovrec V. Clinical outcome of day 2 versus day 5
643 transfer in cycles with one or two developed embryos. *Fertil Steril* 2002;**77**:529–536.

644 Leahy BD, Jang W-D, Yang HY, Struyven R, Wei D, Sun Z, Lee KR, Royston C, Cam L, Kalma Y,
645 *et al.* Automated Measurements of Key Morphological Features of Human Embryos for
646 IVF. *Med Image Comput Comput Assist Interv* 2020;**12265**:25–35.

647 Lee CI, Su YR, Chen CH, Chang TA, Kuo EES, Zheng WL, Hsieh WT, Huang CC, Lee MS, Liu M.
648 End-to-end deep learning for recognition of ploidy status using time-lapse videos. *J*
649 *Assist Reprod Genet* 2021;**38**:1655–1663. Journal of Assisted Reproduction and Genetics.

650 Leung AS, Son WY, Dahan MH. Time-lapse imaging of embryos: current evidence supporting
651 its use. *Expert Rev Med Devices* 2016;**13**:881–883. Taylor & Francis.

652 Liao Q, Zhang Q, Feng X, Huang H, Xu H, Tian B, Liu J, Yu Q, Guo N, Liu Q, *et al.* Development
653 of deep learning algorithms for predicting blastocyst formation and quality by time-
654 lapse monitoring. *Commun Biol* 2021;**4**:415.

655 Lundberg SM, Erion GG, Lee S-I. Consistent Individualized Feature Attribution for Tree
656 Ensembles. 2018;Available from: http://arxiv.org/abs/1802.03888.

657 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf*
658 *Process Syst* 2017;4766–4775.

659 Milewski R, Kuć P, Kuczyńska A, Stankiewicz B, Łukaszuk K, Kuczyński W. A predictive model
660 for blastocyst formation based on morphokinetic parameters in time-lapse monitoring
661 of embryo development. *J Assist Reprod Genet* 2015;**32**:571–579.

662 Mizobe Y, Tokunaga M, Oya N, Iwakiri R, Yoshida N, Sato Y, Onoue N, Ezono Y. Synchrony of
663 the first division as an index of the blastocyst formation rate during embryonic
664 development. *Reprod Med Biol* 2018;**17**:64–70.

Motato Y, los Santos MJ de, Escriba MJ, Ruiz BA, Remohí J, Meseguer M. Morphokinetic analysis and embryonic prediction for blastocyst formation through an integrated time-lapse system. *Fertil Steril* 2016;**105**:376-84.e9.

Raudonis V, Paulauskaite-Taraseviciene A, Sutiene K, Jonaitis D. Towards the automation of early-stage human embryo development detection. *Biomed Eng Online* 2019;**18**:1–20. BioMed Central.

Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: A multicenter study. *Hum Reprod* 2017;**32**:307–314.

Storr A, Venetis CA, Cooke S, Susetio D, Kilani S, Ledger W. Morphokinetic parameters using time-lapse technology and day 5 embryo quality: a prospective cohort study. *J Assist Reprod Genet* 2015;**32**:1151–1160.

Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Process Lett* [Internet] 2014;**21**:1389–1393.

Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time-lapse annotations. *Hum Reprod* 2013;**28**:3215–3221.

the Japan Society of Obstetrics and Gynecology ART Databook 2021Available from: https://www.jsog.or.jp/activity/art/2021_JSOG-ART.pdf.

Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum Reprod* 2019;**34**:1011–1018.

687 Veeck LL. Oocyte Assessment and Biological Performance. *Ann N Y Acad Sci* 1988;**541**:259–
688 274.

689 Vlaisavljević V, Kovačič B, Reljič M, Lovrec VG, Sajko MČ. Is there any benefit from the
690 culture of a single oocyte to a blastocyst-stage embryo in unstimulated cycles? *Hum*
691 *Reprod* 2001;**16**:2379–2383.

692 Wong CC, Loewke KE, Bossert NL, Behr B, Jonge CJ De, Baer TM, Pera RAR. Non-invasive
693 imaging of human embryos before embryonic genome activation predicts development
694 to the blastocyst stage. *Nat Biotechnol* 2010;**28**:1115–1121. Nature Publishing Group.

695 Xiao JS, Healey M, Talmor A, Vollenhoven B. When only one embryo is available, is it better
696 to transfer on Day 3 or to grow on? *Reprod Biomed Online* 2019;**39**:916–923.

697 Yang ST, Shi JX, Gong F, Zhang SP, Lu CF, Tan K, Leng LZ, Hao M, He H, Gu YF, *et al.* Cleavage
698 pattern predicts developmental potential of day 3 human embryos produced by IVF.
699 *Reprod Biomed Online* 2015;**30**:625–634.

700 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;**3**:32–35.

701 Zaninovic N, Nohales M, Zhan Q, los Santos ZMJ de, Sierra J, Rosenwaks Z, Meseguer M. A
702 comparison of morphokinetic markers predicting blastocyst formation and implantation
703 potential from two large clinical data sets. *J Assist Reprod Genet* 2019;**36**:637–646.
704 Journal of Assisted Reproduction and Genetics.

705 Zoph B, Vasudevan V, Shlens J, Le Q V. Learning Transferable Architectures for Scalable
706 Image Recognition. *2018 IEEE/CVF Conf Comput Vis Pattern Recognit* [Internet] 2018;, p.
707 8697–8710. IEEEAvailable from: https://ieeexplore.ieee.org/document/8579005/.

708

709

710 **Figure legends**

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

718

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

724

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

730

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

733

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

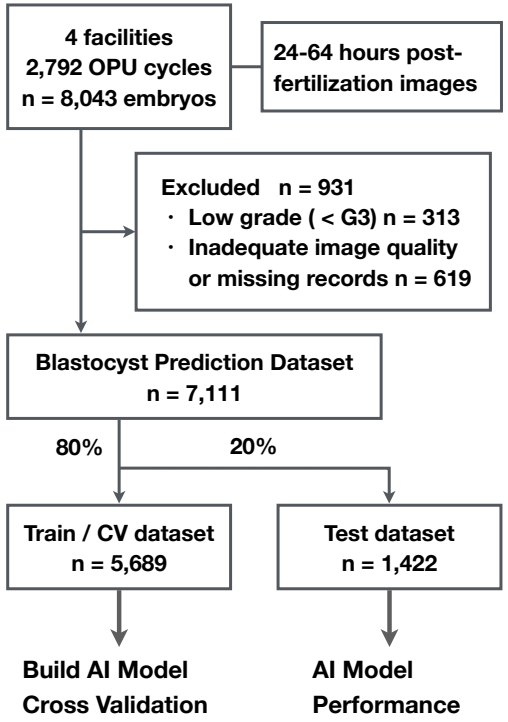
738

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

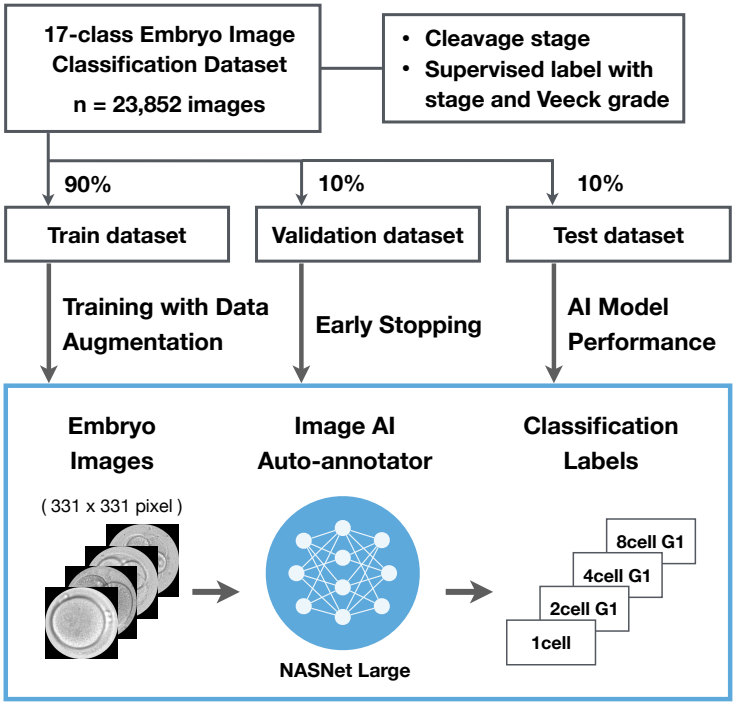
744

Figure 1. Overview of Our Blastocyst Prediction AI

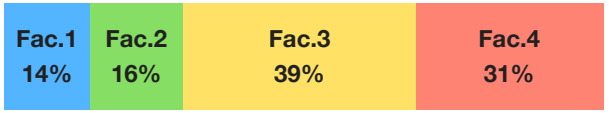
A. Blastocyst Prediction Dataset



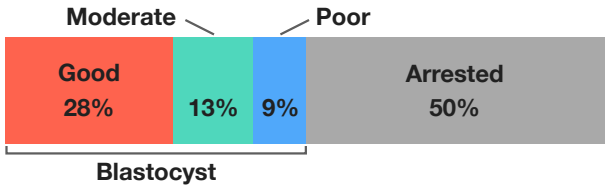
B. Fine-tuning of Auto-annotator



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

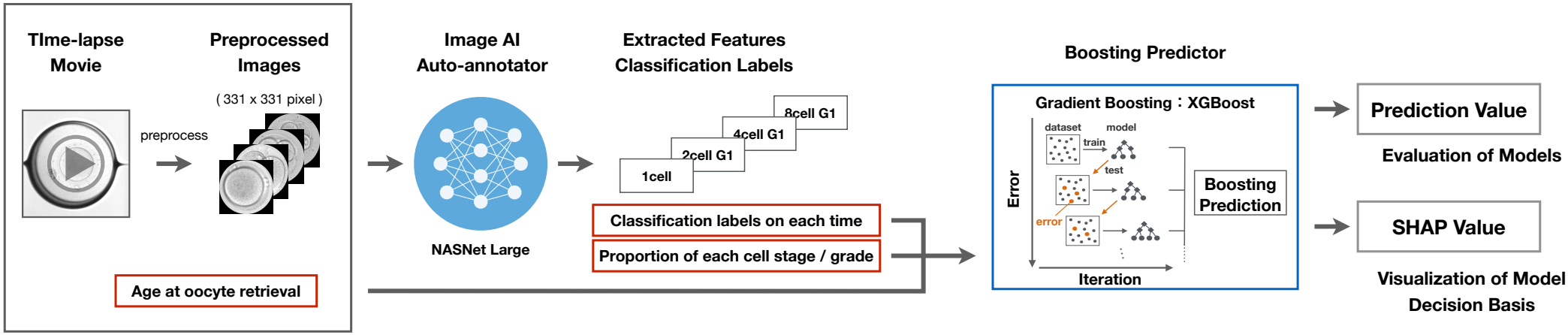
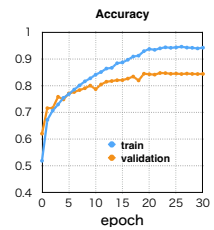


Figure 2. Fine-tuning of Auto-annotator AI

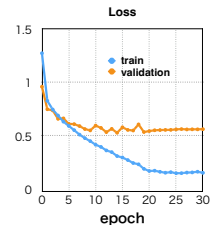
A.

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

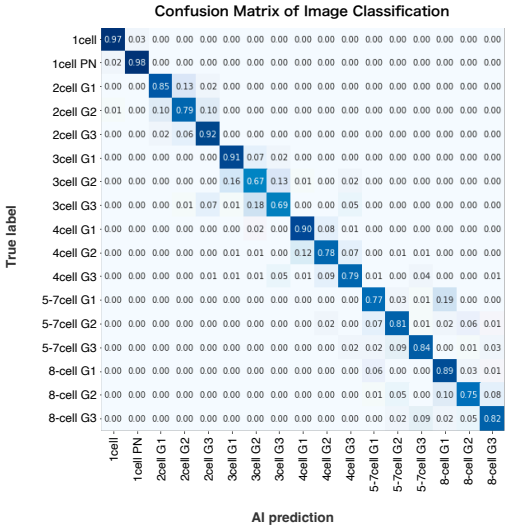
B.



C.



D.



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

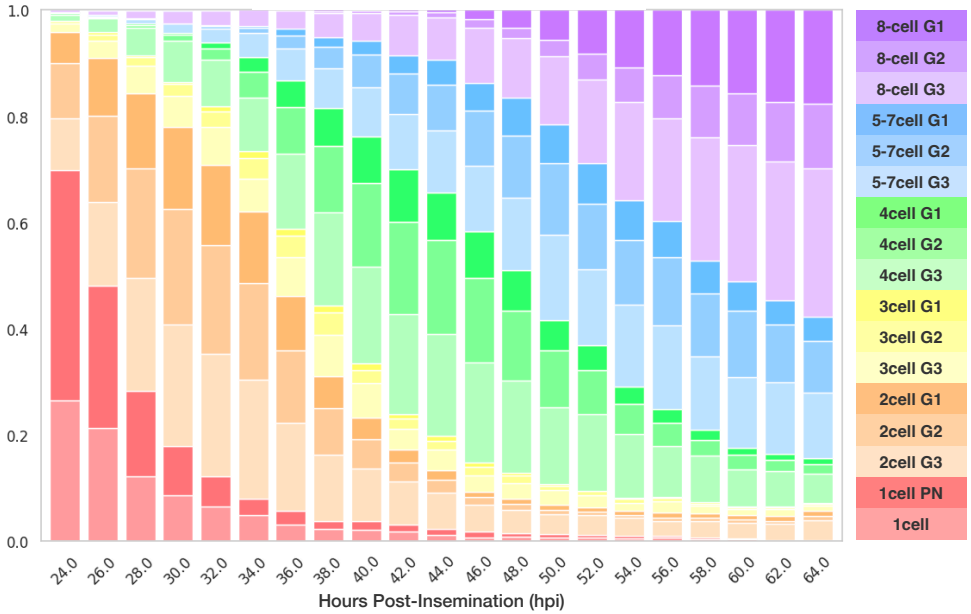
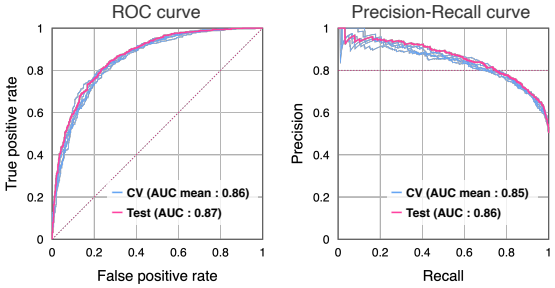
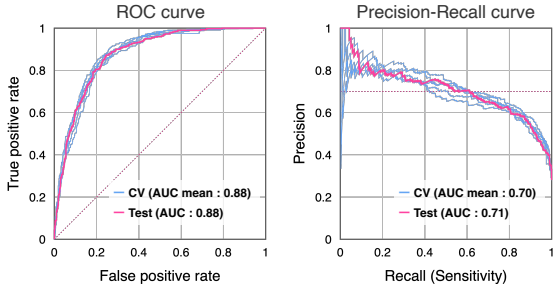


Figure 3. Performance of prediction models

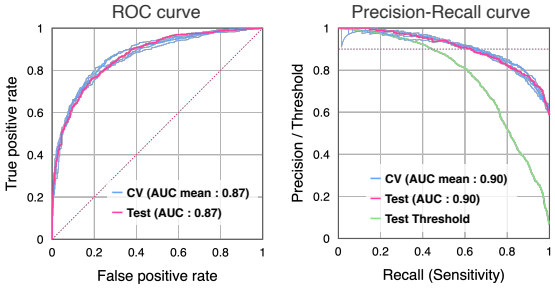
A. Blastocyst formation prediction model



B. Good blastocyst formation prediction model



C. Poor blastocyst + arrested embryo prediction model

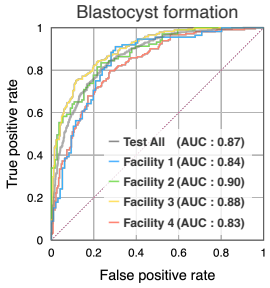


D. Poor blastocyst + arrested embryo prediction model

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	

Figure 4. ROC subgroup analyses

A. Subanalysis by facility



B. Subanalysis by age group

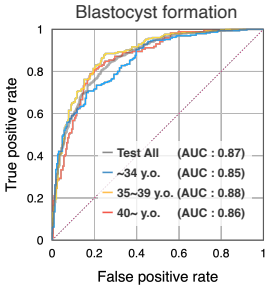
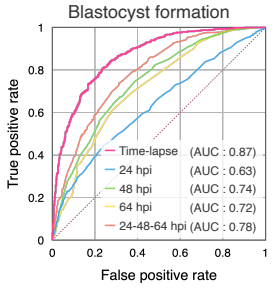


Figure 5.

A. Annotation-based models



B. Image-based models

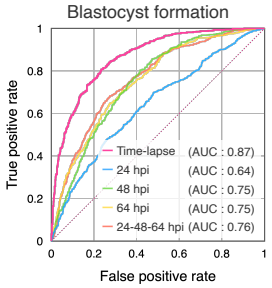
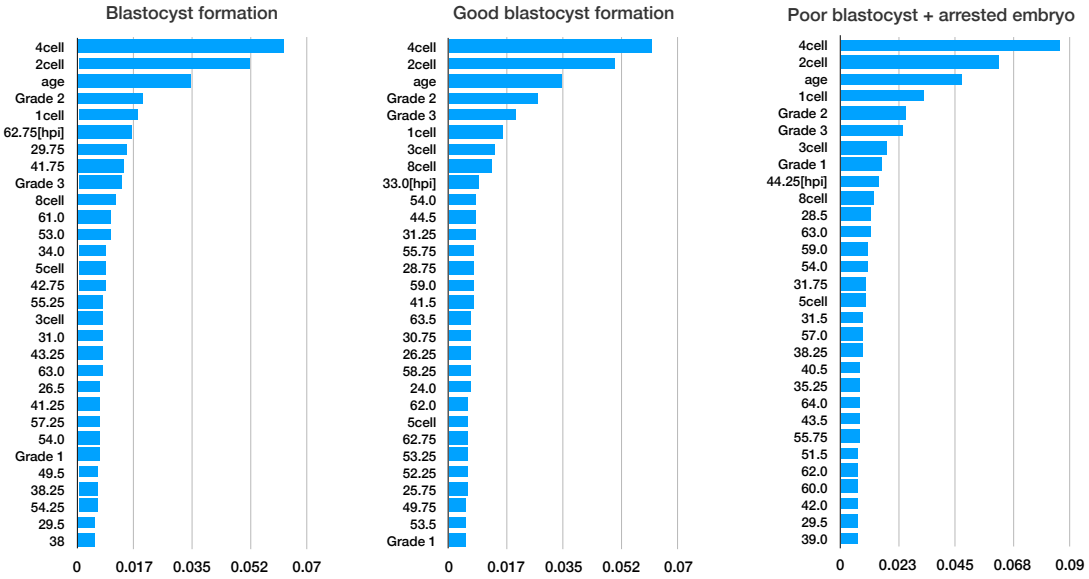
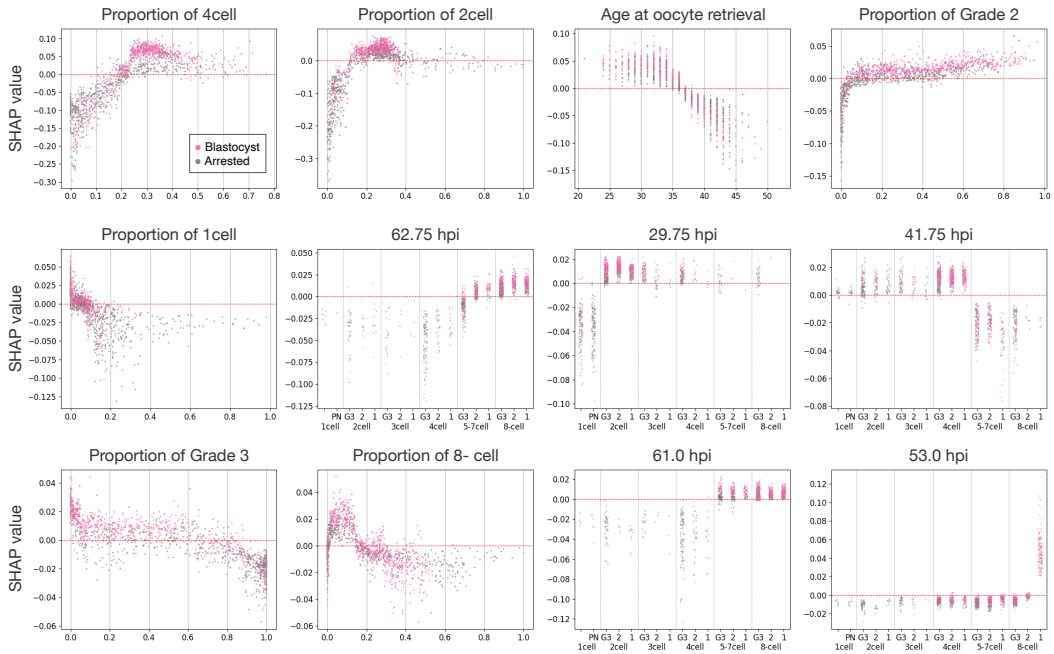


Figure 6. Visualization of model decision basis

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