1   **A novel machine learning based approach for iPS progenitor cell identification**

2

3   Haishan Zhang[1, 2][¶], Ximing Shao[3][¶], Yin Peng[4][¶], Yanning Teng[1, #a], Konda Mani Saravanan[1],

4   Huiling Zhang[1], Hongchang Li[3*], Yanjie Wei[1*]

5

6   [1]Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology,

7   Center for High Performance Computing, Shenzhen Institutes of Advanced Technology,

8   Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China

9   [2]University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District,

10 Beijing 100049, China

11 [3]Shenzhen Key Laboratory for Molecular Biology of Neural Development, Guangdong Key

12 Laboratory of Nanomedicine, Institute of Biomedicine and Biotechnology, Shenzhen

13 Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong

14 518055, China

15 [4]Department of Pathology, Shenzhen University School of Medicine, Shenzhen, Guangdong,

16 PR China 518060

17 [#a]China Merchants Bank Network Technology(Hangzhou) Co., Block B of High-tech

18 Industrial Building, 567 Xincheng Road, Binjiang District, Hangzhou, Zhejiang 310051,

19 China

20

21 [*]Corresponding authors: Hongchang Li - hc.li@siat.ac.cn(HL); Yanjie Wei -

22 yj.wei@siat.ac.cn(YW)

23

24 [¶]These authors contributed equally to this work.

Abstract

Identification of induced pluripotent stem (iPS) progenitor cells, the iPS forming cells in early stage of reprogramming, could provide valuable information for studying the origin and underlying mechanism of iPS cells. However, it is very difficult to identify experimentally since there are no biomarkers known for early progenitor cells, and only about 6 days after reprogramming initiation, iPS cells can be experimentally determined via fluorescent probes. What is more, the ratio of progenitor cells during early reprograming period is below 5%, which is too low to capture experimentally in the early stage.

In this paper, we propose a novel computational approach for the identification of iPS progenitor cells based on machine learning and microscopic image analysis. Firstly, we record the reprogramming process using a live cell imaging system after 48 hours of infection with retroviruses expressing Oct4, Sox2 and Klf4, later iPS progenitor cells and normal murine embryonic fibroblasts (MEFs) within 3 to 5 days after infection are labeled by retrospectively tracing the time-lapse microscopic image. We then calculate 11 types of cell morphological and motion features such as area, speed, etc., and select best time windows for modeling and perform feature selection. Finally, a prediction model using XGBoost is built based on the selected six types of features and best time windows. Our model allows several missing values/frames in the sample datasets, thus it is applicable to a wide range of scenarios.

Cross-validation, holdout validation and independent test experiments showed that the minimum precision is above 52%, that is, the ratio of predicted progenitor cells within 3 to 5 days after viral infection is above 52%. The results also confirmed that the morphology and

2

48   motion pattern of iPS progenitor cells is different from that of normal MEFs, which helps

49   with the machine learning methods for iPS progenitor cell identification.

50

51   ## Keywords

52   iPS progenitor cell; Machine learning; XGBoost; Cell reprogramming; Morphology features

53

54   ## Author Summary

55   Identification of induced pluripotent stem (iPS) progenitor cells could provide valuable

56   information for studying the origin and underlying mechanism of iPS cells. However, it is

57   very difficult to identify experimentally since there are no biomarkers known for early

58   progenitor cells, and only after about 6 days of induction, iPS cells can be experimentally

59   determined via fluorescent probes. What is more, the percentage of the progenitor cells during

60   the early induction period is below 5%, too low to capture experimentally in early stage. In

61   this work, we proposed an approach for the identification of iPS progenitor cells, the iPS

62   forming cells, based on machine learning and microscopic image analysis. The aim is to help

63   biologists to enrich iPS progenitor cells during the early stage of induction, which allows

64   experimentalists to select iPS progenitor cells with much higher probability, and furthermore

65   to study the biomarkers which trigger the reprogramming process.

3

Introduction

Induced pluripotent stem (iPS) cells are cells with embryonic-like state reprogrammed from mouse embryonic or adult fibroblasts by introducing the defined factors[1]. Since Takahashi and Yamanaka[1] first proposed the methods of reprogramming somatic cells to iPS cells, it has become an important method for clinical cell therapy, and revolutionized regenerative medicine[2], such as platelet deficiency[3], spinal cord injury[4], macular degeneration[5], Parkinson's disease[6] and Alzheimer's disease[7]. However, obstacles still remain in scientific and clinical applications for iPS cells because of potential tumorigenicity and low efficiency of reprogramming technique[8-10]. Tumorigenicity is attributed to the introduction of tumorigenic factors such as Oct4, Sox2, Klf4 and c-Myc, of which over-expression is generally associated with tumors. Inefficiency concerns low frequency for reprogramming cells, which is less than a small proportion of 5%. In some induction protocols, the ratio of progenitor cells during the early stage of reprogramming is even under 0.5%.

The above-mentioned obstacles are mainly due to poor understanding of molecular mechanisms in iPS cell reprogramming, which ultimately prevented this technology from a wide range of scientific and clinical applications. Theoretical mechanisms models are proposed such as two-step process model[11] and seesaw model[12], most of which focus on how factors such as Oct4, Sox2, Klf4, and c-Myc induce pluripotency. Experimental approaches based on epigenetic profiling, RNA screening or single-cell analysis for uncovering the mechanisms are limited by the low reprogramming efficiency or the lack of biomarkers for progenitor cells [13-20].

4

89   Recent studies found that iPS progenitor cells differed from normal MEFs in

90   morphology, motion or proliferation rate. Smith et al.[21] found that iPS progenitor cells

91   showed smaller cellular area and higher proliferative rate than normal MEFs via time-lapse

92   imaging. Zhang et al.[22] also found that iPS cells exhibited distinct morphology features and

93   different proliferative rate comparing with larger and quiescent differentiated cells. Li et al.

94   [23] showed the mesenchymal-to-epithelial transition, a process with significant

95   morphological changes, was a key cellular mechanism for induced pluripotency. Megyola et

96   al.[24] demonstrated that migratory motions for progenitor cells were often distinct in

97   direction and distance to bring distant progenitor cells together. Most of these studies relied

98   on time-lapse microscopy, which allowed studying/tracing cellular events in early

99   reprogramming by direct observation [24]. Since iPS progenitor cells exhibit unique

100   morphology and motion features, computational methods, especially machine learning based

101   methods, could provide an alternative method to identify iPS progenitor cells in the early

102   stage of reprogramming process through learning the morphology and motion patterns of iPS

103   progenitor cells.

104   Usually cell detection, segmentation and tracking are firstly required for computational

105   methods to study cell images. Li et al.[25] proposed DCELLIQ for cell nuclei tracking based

106   on neighboring graph and integer programming technique. Dzyubachyk et al.[26] relied on

107   coupled active surfaces algorithm for cell segmentation and tracking in time-lapse

108   fluorescence microscopy images. Maška et al.[27] presented a tracking method for fluorescent

109   cells based on coherence-enhancing diffusion filtering and Chan-Vese model. Türetken et

110   al.[28] proposed an integer programming approach for tracking elliptical cell populations in

111   time-lapse image sequences. Payer et al.[29] developed a recurrent fully convolutional

112   network architecture for instance segmentation and tracking with training network using an

113   embedding loss based on cosine similarities.

114   Recently machine learning/deep learning methods have been extensively developed for

115   the prediction and study of cell images. Using cell images, Erdmann et al.[30] introduced a

116   machine learning based framework for image-based screen analysis. Valen et al.[31] tried to

117   solve cell image segmentation problem utilizing deep convolutional neural networks, and

118   demonstrated its effectiveness in segmenting fluorescent images of cell nuclei. Chen et al.[32]

119   achieved high classification accuracy in label-free white blood T-cells against colon cancer

120   cell via a deep learning method. Similarly with a deep convolutional neural network method,

121   Kraus et al.[33] analyzed the microscopic images for yeast cell and other pheromone-arrested

122   cells, and Gao et al.[34] achieved a high ranking in the human epithelial-2 cell image

123   classification competition hosted by ICPR2014. Together with principal component analysis,

124   machine learning method can be used to infer regulatory network patterns underlying stem

125   cell pluripotency[35]. The ability of machine learning has been demonstrated with its

126   extensive application for cellular image data, however, it has been seldom used in the

127   identification of iPS progenitor cells in the early stage.

128   In this article, we propose a machine learning based approach to detect iPS progenitor

129   cells during the early stage of reprogramming. Given the cell images recorded via live-cell

130   imaging system during the reprogramming process, the paper aims to identify iPS progenitor

131   cells against normal MEFs in the same stage. Since the iPS progenitor cell to normal MEFs

132   ratio is usually below 5%, this makes the identification problem very difficult. In the paper we

6

133   use Imaris, a software from Bitplane, to analyze and process microscopic cell images from

134   live-cell imaging system. Surpass, a module of Imaris is then used to extract cell numerical

135   information in the same time period. We then develop a machine learning method for

136   identification of iPS progenitor cells based on the extracted morphological and motional

137   features. The prediction model is built with XGBoost based on the selected six types of

138   features and time windows. In our method, cell division is not considered, and frames

139   contained in selected time windows are uniform. The model performance is evaluated by

140   three different validation methods. When tested on labeled datasets with a ratio of about 1:5

141   between progenitor cells and normal MEFs, the prediction precision to identify iPS progenitor

142   cells is above 52% during the first 1-3 days of reprogramming after adding iCD1 medium.

143   The image-based machine learning method allows experimentalists to select iPS progenitor

144   cells with much higher probability, and furthermore to study the biomarkers which trigger the

145   reprogramming process.

146

Materials and Methods

148   The workflow used in the paper is presented in **Fig 1**, which mainly includes feature

149   extraction, preprocessing with missing values, feature selection, machine learning for training

150   and validation. In this workflow, we acquire time-lapse images through experiments firstly,

151   then we label iPS progenitor cells and normal MEFs manually to generate datasets by tracing

152   images retrospectively. Next, we generate 11 types of morphology and motion features with

153   Imaris software. After the feature extraction, we perform time window selection and a

154   two-step feature selection. Finally, we build the prediction model based on the selected six

7

155   types of features and six time windows. The machine learning algorithm for modeling is

156   XGBoost, a gradient boosting tree[36]. In the following sections, we will describe the steps of

157   our model in detail.

158   **Fig 1**. **Flow chart of the machine learning based approach for iPS progenitor cell**

159   **identification**

160   In time-lapse imaging, we record the reprogramming process periodically among 54 fields

161   after 48h of viral infection. For retrospective labeling, the figure only shows the labeled cell

162   images of the first frame of all eight phases. Only datasets from phase 1, 2 and 3 are used for

163   model training and testing.

164

165   **Cell culture and generation of iPS cells**

166   Mouse embryonic fibroblasts (MEFs) are derived from E13.5 embryos carrying the Oct4

167   promoter-driven GFP reporter gene[37] and maintain in DMEM (HyClone) supplemented

168   with 10% FBS (Gibco). To generate iPS cells, MEFs within two passages are seeded at a

169   density of 5×10^4 cells/well in 6-well plates and cultured overnight. The next day, MEFs are

170   infected with retroviral supernatants containing the DsRed gene and three reprogramming

171   factors (Oct4, Sox2, Klf4) twice in a 48h process. After 48h of infection, iCD1 medium[38] is

172   changed every day to achieve high reprogramming efficiency. iPS cell colonies are obtained

173   5-7 days post-treatment in iCD1 based on the Oct4-GFP expression.

174

175   **Time-lapse imaging**

176   Reprogramming process is recorded using an Olympus IX81 live cell imaging system

8

177   equipped with a 10× UPlanFL objective, iXon3 EMCCD Camera. The date on which viral

178   supernatants are removed and iCD1 medium are added is defined as Day 0. From Day 0,

179   MEFs images are taken for a total time of 135 hours and 40 minutes. For the first 48 hours

180   and 40 minutes, both bright-field and red fluorescence images are acquired at 10-minute

181   intervals. After two days of the dual-channel imaging, a green fluorescence channel is added

182   to indicate the expression of Oct4-GFP and acquisition intervals are adjusted to 30 minutes.

183   Motorized Stage Control is used to follow cells in the same field and a total of 54 fields are

184   selected at each time for further analysis.

185   Cell images taken within the first 48 hours and 40 minutes since Day 0 are used to

186   construct the dataset because after this time the Oct4-GFP is added to identify the progenitor

187   cells experimentally and the paper tries to identify/predict progenitor cells using

188   computational methods as early as possible.

189

190   **Cell segmentation and numerical feature extraction**

191   The original files are time-lapse microscope images in TIFF format, whose pixels are

192   770 * 746 and the actual size is 1000 microns * 967 microns. Because some fields do not

193   show distinct Oct4-GFP signals and result in no signals for iPS cells in these fields, we only

194   use images from 33 fields for modeling. Imaris (Version 7) software is used to segment cells

195   in the images of these 33 fields and extract the corresponding numerical features for the

196   segmented cells. During this process, the parameter values of cell and nucleus intensity are set

197   the same for all the cells in each field, and cell tracking duration parameter of greater than

198   5000s is used. Imaris utilizes red fluorescent channel for cell segmentation and tracking. The

9

199   image segmentation is based on the Watershed Algorithm, which is very sensitive to weak

200   edges and intensity in images.

201   Features are computed for each segmented/identified cell image at different time frames

202   by Imaris, and these features denote the morphological and movement information of the

203   segmented cells during reprogramming. Overall 11 types of features are extracted (volume,

204   area, sphericity, ellipsoid-prolate, ellipsoid-oblate, nucleus-cytoplasm volume ratio,

205   displacement, speed, Intensity-stdDev, Intensity-Max, Intensity-Min) and each type contains

206   features in several frames of the selected uniform time windows. The detailed list of features

207   is presented in Part 1 of the **S1 File**.

208

209   **Cell image dataset generation**

210   Cell image datasets for machine learning consist of normal MEFs cell images and

211   progenitor cell images within the first 48 hours and 40 minutes. The datasets will be used by

212   our machine learning method in the training and testing processes.

213   At first, we manually label iPS progenitor cell and normal MEFs cell images identified

214   by Imaris software within the first 48 hours and 40 minutes. Experimentally iPS cells can be

215   determined only by Oct4-GFP expression signal, which cannot be observed until the seventh

216   day after transfection with Yamanaka's factors. Cells showing green fluorescence in images

217   are considered as iPS cells. We can then label iPS progenitor cells in the early reprogramming

218   process by cell image backtracking. The corresponding cell images are retrospectively traced

219   frame by frame from GFP expression to the first 48 hours and 40 minutes (**Fig 1**). Due to

220   three one-hour iCD1 medium changes, the total reprogramming period is divided into four

221   periods, the first period is 16 hours and 50 minutes long, from 18 hours to 24 hours and 40

222   minutes denoted as phase 1 in the paper, the second from 25 hours and 50 minutes to 40 hours

223   and 40 minutes denoted as phase 2, and the third from 41 hours and 50 minutes to 48 hours

224   and 40 minutes denoted as phase 3. In this paper, we focus on these three periods (phases 1, 2

225   and 3) only because of tiny ratio for iPS progenitor cells in the first 16 hours and 50 minutes,

226   which is even less than 2%.

227   Two rules are applied in the paper for generating the cell image datasets, (1) cell division

228   is not considered; (2) frames from the same window of each phase are selected for modeling

229   among uniform time periods. When cell division is taken into account, features in the mother

230   cell and its daughter cells are not comparable, for example, the area of mother cell is much

231   bigger than that of its daughter cells, thus the machine learning model will fail to process this

232   cell. The second rule guarantees that time dimension (time period and length) for the cell

233   image data samples should be uniform.

234   For each cell, not every image in different frames can be identified by Imaris due to the

235   fact that different parameter settings (cell or nucleus intensity threshold, cell tracking duration)

236   by Imaris will lead to different segmented cell images in a frame. This results in cell image

237   data missing in some frames, thus our method allows a certain number of missing cell images

238   in the selected uniform time periods and tries to find the maximum number of continuous

239   cells images in this uniform time period.

240   Overall three cell image sets are generated for three phases, each with an approximately

241   1:5 ratio between progenitor cell images and normal MEFs cell images. For phase 1, 78 IPS

242   progenitor cells and 391 normal MEFs are labeled; for phase 2, 84 IPS progenitor cells and

11

243 420 normal MEFs are labeled; for phase 3, 74 IPS progenitor cells and 370 normal MEFs are

244 labeled. Each of these three initial cell image sets are divided into the training and test sets:

245 70% of cell images for each time phases are selected randomly as training set with the

246 remainder (30%) as test set. The ratio between progenitor cell images and normal MEFs cell

247 images is kept approximately 1:5 for these training and testing sets. For the the training sets,

248 there are 55 iPS progenitor cells and 274 normal MEF cells in phase 1, 59 iPS progenitor cells

249 and 294 normal MEF cells in phase 2, as well as 52 iPS progenitor cells and 259 normal MEF

250 cells in phase 3.

251 In this paper, the initial cell dataset is used for cross-validating the proposed method, and

252 the training dataset is used for missing value processing and feature selection. For different

253 analytic steps, the specific data sample size depends on the time period from which the data

254 has been collected. Numerical features are calculated for all cell images in the datasets and

255 saved in CSV files. All datasets are standardized utilizing z-score.

256

257 **Missing values processing**

258 Processing missing values for the cells in the corresponding frames is an important step

259 for our model. Imaris cannot continuously identify all the cells in the frame due to different

260 parameter settings or complex three-dimensional cell environment. This implies that there

261 exists a certain number of cell images with missing feature values in the uniform time periods.

262 A certain number of missing images in the frames are permitted for cells to guarantee a

263 modest data size, and missing cell features are estimated with an imputation method. To

264 choose the most appropriate approach, we first analyze the impact of the number of missing

12

265   frames on the model, and then analyze the effect of three different imputation methods under

266   the corresponding missing frame numbers. Details for the three imputation methods are as

267   follows:

268   • *set_mean*. The missing value is set to the average value of all nonempty frames for a

269   specific type of feature in its sample from the selected time window.

270   • *set_KNN*. The missing value is set to the weighted average value of five nearest

271   nonempty neighbor frames for a specific type of feature in its sample. The calculation

272   of weight uses k-Nearest Neighbor (KNN) algorithm. The formula is as

$$\text{Missing value}_{\text{frame}_i} = \sum_j w_{frame_j} \cdot feature_{frame_j}, \ w_{frame_j} = \frac{1}{\sqrt{(i-j)^2}} \quad (1)$$

273

274   where j represents the index of five nearest frames neighbor for missing frame i.

275   • *set_mean_mod*. Missing value is set to the average value of five nearest nonempty

276   neighbor frames for a specific type of feature in its sample.

277

278   **Time window and feature selection**

279   Because of the two rules used in dataset generation (Section **Cell image datasets**

280   **generation**), although images are provided up to 49 hours, it is unable to construct the model

281   based on the whole period. From a total of 49 hours, numerous time periods can be chosen,

282   and the model needs to select best time windows among all these eligible time periods. Time

283   window selection includes start frame selection and window length selection. Start frame

284   represents the moment that the time window starts from, and window length represents frame

285   number that the time window contains. For each time window with a selected time frame and

286   window length, we train and validate the proposed method on the corresponding dataset

13

287   generated. Validation is performed with 5-fold cross validation and the evaluation metric is

288   precision.

289   Morphological and motion feature selection is used to improve the performance. Since it

290   is difficult to guarantee image recording time to be accurately consistent for every batch

291   through experiments, model performance needs to be robust among wider time periods. Every

292   type of features contains multiple frames of features from the corresponding best time

293   windows. Features in a time window are treated as a bundle so we can learn the dynamic cell

294   growth process.

295   There are two steps for feature selection. The first step is recursive feature elimination.

296   Firstly, we use all 11 types of features to train the model with 5-fold cross validation and

297   calculate its precision as initial unimportance score. Then we delete each type of feature at a

298   time and obtain 11 precision values as new unimportance scores. We compare every new

299   score with the initial score, and remove the feature type with the largest unimportance score

300   higher than initial score. The recursive process will be repeated on feature set until the model

301   performance can be no longer improved or there is no feature. We then rank the importance of

302   all 11 types of features and delete the least important feature types. Second, we calculate the

303   Pearson correlation coefficient for the selected feature types from step 1 to remove the highly

304   correlated features with a correlation coefficient of 0.60 or above.

305

306   **Machine learning model and validation**

307   XGBoost, a Boosting algorithm, is used in this paper for feature selection and IPS cell

308   recognition. XGBoost integrates many weak tree-classifiers together to form a strong

14

309   classifier. This algorithm applies numerous strategies to prevent overfitting, and it is widely

310   utilized in data science such as cell analysis [39-43]. Hyperparameters of XGBoost are tuned

311   using grid-search for model training with selected features and best time windows.

312   For model validation, firstly we use 5-fold cross-validation on the initial cell image

313   datasets from the time windows of the three phases. Dataset generated from initial cell-sets

314   contains about 70 iPS cells for each phase. The ratio of iPS cells and normal MEFs keeps as

315   1:5 in each dataset.

316   In order to test the model's ability/robustness to predict the iPS progenitor cells around

317   the neighborhood of the corresponding training time window, holdout validation is performed.

318   Because iCD1 medium change is operated manually during the experiments, it is

319   impracticable to guarantee that for per batch data the duration of medium change is accurately

320   consistent with the existing data. This inconsistency might lead to a non-exact match between

321   the timeline after medium change and the timeline used in the model training process. The

322   holdout validation is designed as follows, for the model trained on time window $i \sim j$, we

323   examine the model's performance on several neighbor time windows, including time windows

324    $i\text{-}3 \sim j\text{-}3, i\text{-}2 \sim j\text{-}2, i\text{-}1 \sim j\text{-}1, i \sim j, i\text{+}1 \sim j\text{+}1, i\text{+}2 \sim j\text{+}2,$ and $i\text{+}3 \sim j\text{+}3$, where i represents start

325   time frame of the window and j represents the terminal frame. The training dataset from time

326   window $i \sim j$ is generated from the initial training image data sets (70% of the initial total

327   dataset), and test datasets of the seven neighbor time windows are generated from the test

328   datasets (30% of the initial total dataset).

329   Moreover, in order to further test our model's ability to predict the iPS progenitor cell on

330   a time window which doesn't overlap with the window in the training process, an independent

15

331   test is performed. Model performance is tested on time windows which are far away from the

332   training time windows. Since we have three time phases, we first select test time windows in

333   phase 2 and 3 for the models trained on time windows of phase 1 and 2 respectively. For

334   testing our model developed for phase 3, we select the independent test time windows also in

335   phase 3, but without any overlap with the corresponding training time windows.

336

337   **Evaluation metrics**

338   In this paper, precision is mainly used for evaluation defined as,

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

339

340   where TP and FP represent the number of true positive and false positive prediction. This

341   metric evaluates the accuracy for the positive sample predicted by the model. Biologists need

342   a cell sample set enriched with true iPS progenitor cells so that in the early stage of

343   reprogramming progenitor cells can be studied with high probability.

344

345   **Results and Discussion**

346

347   **Missing frames processing and imputation method**

348   First, the effect of missing frames and imputation methods on the model's performance

349   was analyzed. Experiment for missing value was performed under six kinds of missing frame

350   numbers, which were numbers below or equal to five, four, three, two, one and zero. Model

351   performance was tested for each missing frame number with three imputation methods on

352   time periods of two window lengths (10 and 19 frames) located in three phases, which were

16

353   time period/window 19h30min ~ 21h10min from phase 1 (TP1), 25h50min ~ 27h30min from

354   phase 1 (TP2), 41h50min ~ 43h30min from phase 2 (TP3), 18h10min ~ 21h20min from phase

355   2 (TP4), 26h ~ 29h10min from phase 3 (TP5) and 42h ~ 45h10min from phase 3 (TP6).

356   Two window lengths (10 and 19 frames) were selected because a reasonable number of

357   continuous cell images could be traced. A short window will have more data but the motion

358   and morphological pattern of iPS progenitor cells cannot be learned while a long window will

359   result in a much smaller dataset. For each length, we chose three time windows randomly to

360   study whether different lengths would affect model performance under uniform missing frame

361   number. Datasets were generated from the training datasets, which were about 52~59 iPS

362   cells and 259~294 normal MEFs for time windows with 10 frames, 43~50 iPS cells and

363   238~264 normal MEFs for time windows with 19 frames. Model was evaluated by the

364   average precision with 5-fold cross validation over 20 times.

365   **Fig 2** showed the comparison results of different missing frame numbers and imputation

366   methods. For each missing number and imputation method, **Fig 2(a)** described the average

367   precision over six time windows (TP1 to TP6), indicated by blue boxes for set_KNN, red

368   boxes for set_mean and green boxes for set_mean_mod. Also shown in **Fig 2(a)** was the

369   average precision over all three imputation methods, indicated by grey boxes. **Fig 2(b)**

370   described the standard deviations of the corresponding precision values in **Fig 2(a)**. Detailed

371   precisions for all six time periods (TP1~TP6) were provided in Figure S1 of the **S1 File**.

372   **Fig 2. Model comparison for different missing frame number and imputation methods**

373   Fig 2(a) shows the average precision over six time periods (TP1 to TP6) for each missing

374   frame number and imputation method set_KNN (colored as blue), set_mean (colored as red),

17

375   set_mean_mod (colored as green) and all three imputation methods (colored as gray). Fig 2(b)

376   shows the standard deviation, as a function of missing frame number, of imputation method

377   set_KNN (colored as blue), set_mean (colored as red), set_mean_mod (colored as green) and

378   all three imputation methods (colored as gray).

379   **Fig 2(a)** showed that precision was higher when several missing frames were allowed.

380   For missing frame number of 0, the average precision of all method was only 0.585 and all

381   the average precisions of non-zero missing frame numbers were higher than 0.585. **Fig 2(a)**

382   also showed that the maximum average precision of all method was about 0.632 under

383   missing frames of 4, 4.7% higher than precision under no missing frames and 0.9% higher

384   than precision under missing frame number of 2. On one hand, the size of the dataset is larger

385   when missing value is permitted, on the other hand, the missing frame may introduce new

386   pattern for classification because iPS progenitor cells proliferate more frequently than normal

387   MEFs, and cell division can partly result in missing value. When cells divide at a certain

388   frame in their time periods, the feature values of all subsequent frames are missing.

389   In **Fig 2(b)**, the maximum standard deviation of all methods as indicated by gray box

390   was 0.061 under missing 4 frames. For each specific method, the maximum standard

391   deviation was 0.081 for Set_mean under 5 missing frames. The precision with two missing

392   frame numbers had the minimum standard deviation for all method (0.048 as indicated by

393   gray boxes) and at the same time it was also very close to the maximum precision (0.623

394   compared with the maximum value of 0.632 in **Fig 2(a)**). In addition, Set_mean_mod showed

395   the minimum standard deviation of all 3 imputation methods for all missing frame numbers

396   (indicated by green boxes), an indication of stable performance. Although Set_mean_mod

18

397 also showed smallest standard deviation for missing frame number of 1, its precision value of

398 missing frame number was smaller than that of missing frame number of 2. Therefore, we

399 used missing frame number less than or equal to two and select imputation method as

400 set_mean_mod in our model.

401

402 **Time window selection**

403   Time window selection was performed to select best time windows with high precision

404 for each phase. Since Imaris could not detect all cell images in every frame, the whole time

405 periods of three phases were divided into numerous time windows. For time window selection

406 (including start frame and window length), we set start frame to 21 time points which were

407 18h20min, 18h40min, 19h, 19h20min, 19h40min, 20h, 20h20min, 26h10min, 26h30min,

408 26h50min, 27h10min, 27h30min, 27h50min, 28h10min, 42h10min, 42h30min, 42h50min,

409 43h10min, 43h30min, 43h50min, 44h10min in three phases. Meanwhile, we set window

410 length to 12 different values including 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27 and 29 frames.

411 For the total of 252 (12 times 21) time windows, we first generated datasets for each time

412 window with 11 types of morphological/motion features. All datasets were generated based

413 on the training dataset and contained about 38~59 iPS progenitor cells and about 190~295

414 normal MEFs. Then we selected the optimal time window through 5-fold cross-validation

415 based on 20 XGBoost runs.

416   The model performance on these different time windows was shown in **Fig 3**. In this

417 figure, shorter window lengths were marked in red colors and longer window lengths were

418 marked in blue colors. We observed that precision of longer window lengths was lower than

19

419   that of shorter window lengths in three phases, and this trend was less pronounced for phase 3.

420   The size of the dataset may be the major reason for this trend. Due to the two rules in dataset

421   generation, the amount of samples satisfying conditions decreases gradually with the

422   increasing window length. For window length of 29 frames, there are just about 38 iPS

423   progenitor cells and 262 normal MEFs in phase 1, about 44 iPS progenitor cells and 283

424   normal MEFs in phase 2, about 36 iPS progenitor cells and 242 normal MEFs in phase 3. As

425   compared with the window length of 7 frames, there are about 53 iPS progenitor cells and 290

426   normal MEFs in phase 1, about 55 iPS progenitor cells and 285 normal MEFs in phase 2,

427   about 57 iPS progenitor cells and 300 normal MEFs in phase 3. On the other hand, the

428   number of samples is much less for later start frame than that for previous time since some

429   cells have divided. For instance, there are only about 30 iPS progenitor cells and 200 normal

430   MEFs for the last start frame with length of 29 frames in phase 3.

431   **Fig 3. Time window selection**

432   The three subplots represent the precision values for different time windows based on 21 start

433   frames (x axis) and 12 window lengths (7 frames to 29 frames) for phases 1, 2, and 3 (from

434   top to bottom) respectively, and the black bash line in each subplot indicates a precision value

435   of 0.55.

436   Selection of best time windows according to maximum precision resulted in an unstable

437   prediction performance. For instance, precision achieved the maximum value on the time

438   window starting at 43h30min with length of 29 frames while all its adjacent time windows

439   had poor performance with lower precision. It is unlikely to achieve the same performance on

440   a new dataset of the same time window.

20

441   We selected the best start frame for each phase respectively. To exclude the start frame

442   with high prediction precision for only 1 or 2 window lengths, 14 candidates of best start

443   frames were selected when precision was above 0.55 for at least three successive window

444   lengths. For each candidate best start frame, the average precision was calculated over the

445   successive window lengths whose precision was above 0.55 and the average precision values

446   were shown above each candidate best start frame in **Fig 3**. We only selected one best start

447   frame for each phase according to the average precision values of the candidate best start

448   frames, resulted in 19h40min, 26h10min and 42h30min for phases 1, 2 and 3, respectively.

449   Secondly, the candidate best window lengths were selected whose precision values were

450   all above 0.55 for 3 best start frames of step 1, resulting in window lengths 11, 13, 15 and 17

451   frames. For each window length, the precision values, average precisions and the

452   corresponding standard deviation of 3 different best start frames were provided in Table S1 of

453   **S1 File**. The average precision of 0.640 for window length of 13 frame was the highest while

454   its standard deviation was the smallest (0.01), thus window length of 13 frames was selected

455   as the best window length.

456

457   **Two-step feature selection**

458   We performed a two-step feature selection method on three phases respectively. Firstly,

459   we generated datasets from best time windows based on the training cell image datasets. The

460   dataset of each phase contained 11 types of morphological and motion features, all of which

461   contained about 50~59 iPS progenitor cells and about 200~295 normal MEFs.

462   For the first step, an iterative feature removal procedure was performed on the

21

463   corresponding dataset of each phase to study the importance of each feature type. Average

464   precision was calculated via 5-fold cross-validation over 20 runs on the dataset of each phase,

465   and later sets as initial unimportance score. Next, we removed each type of features and

466   calculated the unimportance scores (average precision). Feature with maximum score would

467   be deleted only if this score was greater than the initial unimportance score, which would then

468   be updated as the maximum score. This step was repeated until no score was greater than

469   initial score or no more feature could be selected.

470   Results from step 1 feature selection were shown in **Fig 4**. For phase 1 precision was no

471   longer improving after removing ellipsoid-oblate, displacement and volume; for phase 2

472   precision was no longer improving after removing displacement and volume; for phase 3

473   precision was no longer improving after removing displacement, ellipsoid-prolate, area and

474   volume. In the end, eight types of features were selected for phase 1, nine types of features

475   were retained for phase 2, and seven types of features were retained for phase 3. Selected

476   features from this step were indicated in **Fig 4** by star symbols. The corresponding precisions

477   for best windows with 13 frames before feature selection were 0.624, 0.607, 0.646 for phases

478   1, 2 and 3, respectively, and after feature selection, these precision values had increased to

479   0.691, 0.613 and 0.682 respectively.

480   **Fig 4. Feature ranking and selection**

481   This figure shows how the precision values change with the deleted feature in a recursive

482   fashion. Least important features are removed earlier.

483   The removing order of feature type in **Fig 4** indicated the importance of each feature

484   type. We observed from **Fig 4** that three types of features, nucleus-cytoplasm ratio, sphericity

22

485   and intensity-StdDev, were important among all three phases. Nucleus-cytoplasm ratio was

486   the top important factor in three phases. Sphericity and intensity-StdDev were among the top

487   4 common features of three phases. Intensity showed clear different patterns between normal

488   MEFs and progenitor cells. As shown in **Fig 5(a)**, the progenitor cells in the blue circles

489   showed a uniform intensity distribution between nucleus and cytoplasm, while for normal

490   MEFs in the yellow boxes, the cytoplasm showed weaker intensity as indicated by the

491   blurring edges. Also shown in **Fig 5(a)**, the nucleus and cytoplasm of progenitor cells in the

492   blue circles and normal MEFs in the yellow boxes were enlarged and colored by light blue

493   and green respectively. It is clear that nucleus-cytoplasm ratio for progenitor cells are much

494   larger than that of normal MEFs. From **Fig 5(a)**, the cell area of progenitor cells is also

495   smaller on average than normal MEFs, indicating the importance of sphericity since area is

496   closely related to sphericity by the equation from Part 1 of the **S1 File**. The selected features

497   are consistent with the experimental results that iPS progenitor cells exhibit higher

498   nucleus-cytoplasm ratio, smaller total area, and higher proliferation rate than normal

499   MEFs[21].

500   **Fig 5. iPS progenitor cells vs. MEFs and Feature correlation**

501   (a) shows the examples of iPS progenitor cell images (blue circles) and normal MEFs images

502   (yellow boxes) taken from phase 1, 2 and 3 of field 2 (Left, middle and right). Nucleus and

503   cytoplasm of the enlarged progenitor cells and normal MEFs are colored in light blue and

504   green respectively. (b) shows the Pearson coefficients between remaining types of features in

505   three phases after the first step of feature selection. Note in this figure ellipsoid-prolate is

506   denoted as E-prolate, intensity-StdDev as I-stdDev, intensity-min as I-Min, intensity-max as

23

507   I-Max, nucleus-cytoplasm volume ratio as Ratio, ellipsoid-oblate as E-oblate.

508   In order to further study the correlations of different features, as a second step we

509   calculated the Pearson correlation coefficients between the selected features. The results for

510   three phases were shown in **Fig 5(b)**. In our model, two feature types were considered

511   strongly correlated if the coefficient was greater than 0.6 and one of them was removed.

512   When two different feature types were strongly correlated with a third feature type, both of

513   them were removed with the purpose of keeping as less number of features as possible. For

514   phase 1, the coefficient between sphericity and area was 0.77 in phase 1, and the coefficient

515   between sphericity and ellipsoid-prolate was 0.66, thus area and ellipsoid-prolate were

516   removed from the list. Similarly, they were removed for phase 2 as well. The strong

517   correlation between sphericity, ellipsoid-prolate and area is caused by the fact that Imaris

518   extracts features from two-dimensional cell images assuming cell thickness as constant.

519   Furthermore, since ellipsoid-oblate was associated with cell thickness, it was removed from

520   the feature list as well for phase 2 and phase 3. Overall, six types of features (Sphericity,

521   I-Min, I-stdDev, I-Max, Ratio, Speed) were selected for all the models.

522

523   **Cross-validation**

524   With selected features, a grid-search scheme was used for hyperparameter optimization

525   of XGBoost with 5-fold cross-validation, and the datasets were generated based on the

526   training sets for three phases. Three hyperparameters such as learning_rate, n_estimators and

527   gamma were set to 0.01, 385 and 0 respectively. We had validated our model with three

528   different experiments as shown in **Fig 1**.

24

529   For cross-validation, datasets were generated from initial whole cell image dataset.

530   Dataset for phase 1 contained about 63 iPS progenitor cells and about 326 normal MEFs.

531   Dataset for phase 2 contained about 82 iPS progenitor cells and about 427 normal MEFs.

532   Dataset for phase 3 contained about 72 iPS progenitor cells and about 359 normal MEFs. For

533   each phase, 5-fold cross validation was performed 10 times on every best time windows with

534   6 selected feature types, resulting in a total of 117 for window length of 13 frames. **Fig 6(a)**

535   showed precision scores for 3 different phases, and all of the precision values were above

536   0.580. For phase 1, the precision value was highest, 0.732.

537   **Fig 6 Model validation**

538   In all sub-figures, X axis indicates the start frame of the best time windows and the

539   corresponding window length (13 frames) is indicated in the inlet. (a) 5-fold cross-validation

540   precisions over 10 runs. (b) the standard deviation of the average precision of the

541   neighborhood time windows in Figure 6(d). (c) the standard deviation of the average

542   precision of the distant windows in Figure 6(e). (d) the average precision of seven

543   neighborhood time windows calculated over 10 holdout validation runs. (e) the average

544   precision over 10 independent tests for six best time windows on their corresponding distant

545   windows.

546

547   **Holdout validation**

548   Holdout validation was used to test the model's ability to predict the iPS progenitor cells

549   in the neighborhood of the time window in which the model had been trained. Since in real

550   application, it is difficult to generate the dataset whose images have the exact start time as in

25

551 the training dataset, holdout-validation is very important for testing the model's generality on

552 the neighborhood time windows. For each phase, the training dataset for window length of 13

553 frames was generated. In phase 1, the window start frame I was 19h40min as shown in **Fig**

554 **6(d).** Models trained on this dataset was then tested on seven test datasets corresponding to

555 start frames I, I-1, I-2, I-3, I+1, I+2 and I+3, illustrated in **Fig 1** and **Fig 6(d).** There was no

556 overlap between the training and testing datasets.

557 For each time window, average precision value was computed over 10 holdout validation

558 runs, and the results were shown in **Fig 6(d).** The minimum average precision values were

559 0.616 for window length of 13 frames and start frame I-2 in phase 1, 0.522 for window length

560 of 13 frames and start frame I-2 in phase 2 and 0.566 for window length of 13 frames and

561 start frame I-3 in phase 3. These minimum precisions were all smaller than the corresponding

562 precisions in **Fig 6(a)**; what is more, **Fig 6(d)** also showed the average precision values for

563 phase 1, 2 and 3 were all smaller than the cross-validation resulted in **Fig 6(a)**, indicating the

564 difficulties for predicting the neighborhood time windows.

565 For each result of the 3 phases in **Fig 6(d),** the standard deviations of average precisions

566 were computed for window length of 13 frames in **Fig 6(b)**. The maximum deviation was

567 0.042 for window length of 13 frames in phase 1 and this indicated the trained models were

568 relatively stable in terms of prediction precision in a wide range of neighborhood windows.

569

570 **Independent test**

571 Finally, to test the model's ability to predict the iPS progenitor cells on a distant time

572 window without overlapped frames with the training window, we performed an independent

26

573 test. If the training cell trajectory is long and contains enough typical iPS progenitor cells, the

574 trained model on one window should be able to identify the motion and morphological

575 patterns of iPS progenitor cells against normal MEFs, regardless of the selected time window.

576 For phase 1, the model trained on time window 19h40min~21h40min (length of 13

577 frames) was tested on time windows of phase 2, including time windows starting from

578 26h20min (S11), 26h40min (S12), 27h (S13), 27h20min (S14), 27h40min (S15), and 28h

579 (S16), shown in the first panel of **Fig 6(e)**. Similarly, for phase 2, the model trained on time

580 windows 26h10min~28h10min (length of 13 frames) was tested on six time windows of

581 phase 3 starting from 42h10min (S21), 42h30min (S22), 42h50min (S23), 43h10min (S24),

582 43h30min (S25), 43h50min (S26), shown in the middle panel of **Fig 6(e)**. Lastly, for phase 3,

583 model testing was performed on the distant time windows without overlapped frames from

584 the same phase, shown in the right panel of **Fig 6(e).** For time windows 42h30min~44h30min,

585 we selected test time windows starting from 45h10min (S31), 45h30min (S32), 45h50min

586 (S33), 46h10min (S34), 46h30min (S35).

587 Results of the independent test runs were shown in **Fig 6(e)**. The minimum precision was

588 0.523 for window length of 13 frames for S16 in phase 1. The average precision of phase 1

589 was lower than those of holdout validation and cross-validation, however, the average

590 precision of phase 2 and 3 were both better than cross-validation and holdout validation. For

591 the prediction of distant time windows, our model could have worse performance than that of

592 neighborhood windows, but our model could also outperform the cross validation and holdout

593 validation (indicated by the standard deviation in **Fig 6(c)**). The reason was the independent

594 test datasets for phase 2 and 3 were closely related to the training dataset. The standard

27

595   deviations of the independent tests were much higher than those of the holdout validation,

596   which could also be seen from the large fluctuations of the precision values in **Fig 6(e)**.

597   Nevertheless, the minimum average prediction precision was above 52% among all the

598   experiments, and maximum average precision was about 0.750 for the independent test in

599   phase 3.

600

Conclusion

602   In this paper, we proposed a machine learning based model together with time-lapse

603   image analysis to predict/identify iPS progenitor cells during the first 3-5 days after

604   reprogramming initiation. The model generated a variety of morphological and motion

605   features among different time windows, then relied on a two-step feature selection algorithm

606   to select the most important features. The proposed computational approach is very unique

607   from previous experimental techniques which identify the iPS progenitor cells by

608   retrospectively tracking the cell images manually frame by frame from the image frame of

609   GFP expression.

610   By the experimental study of the enriched iPS progenitor cells in the early stage of

611   reprogramming, the proposed method could provide a new technique or attempt for

612   experimenters to improve the iPS reprogramming efficiency and to study the underlying

613   mechanism of iPS reprogramming. Morphological and motion features, especially sphericity,

614   intensity-StdDev and nucleus-cytoplasm volume ratio, have been found most important for

615   the progenitor cell classification, which is consistent with the experimental observations.

616   Cross-validation of the proposed method trained and tested on the same time window

28

617 showed that the prediction precision is above 0.580 for all three phases. Since in real

618 applications, it is very difficult to match imaging timeline precisely between different

619 experiments, holdout validation and an independent test are also performed to test the model's

620 ability to predict iPS progenitor cells in the neighborhood time windows and distant time

621 windows, respectively. The results showed our model can predict the iPS progenitor cells

622 with a minimum precision of 52% for neighborhood windows and distant windows, and the

623 maximum average precision is about 0.750 for the independent test in phase 3. The prediction

624 performance of our model tends to have a larger fluctuation for distant windows than for

625 neighborhood windows, indicated by the larger standard deviation of independent test runs.

626 For future works, models on different time windows for each phase can be combined to

627 achieve higher prediction accuracy.

628

Acknowledgment

638

29

Reference

640

641   1.   Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult
642   fibroblast cultures by defined factors. Cell. 2006;126(4):663-76.

643   2.   Yamanaka S. Induced pluripotent stem cells: past, present, and future. Cell Stem Cell.
644   2012;10(6):678-84.

645   3.   Takayama N, Nishimura S, Nakamura S, Shimizu T, Ohnishi R, Endo H, et al. Transient
646   activation of c-MYC expression is critical for efficient platelet generation from human induced
647   pluripotent stem cells. J Exp Med. 2010;207(13):2817-30.

648   4.   Nori S, Okada Y, Yasuda A, Tsuji O, Takahashi Y, Kobayashi Y, et al. Grafted human-induced
649   pluripotent stem-cell-derived neurospheres promote motor functional recovery after spinal cord injury
650   in mice. Proc Natl Acad Sci U S A. 2011;108(40):16825-30.

651   5.   Okamoto S, Takahashi M. Induction of retinal pigment epithelial cells from monkey iPS cells.
652   Invest Ophthalmol Vis Sci. 2011;52(12):8785-90.

653   6.   Kriks S, Shim JW, Piao J, Ganat YM, Wakeman DR, Xie Z, et al. Dopamine neurons derived
654   from human ES cells efficiently engraft in animal models of Parkinson's disease. Nature.
655   2011;480(7378):547-51.

656   7.   Israel MA, Yuan SH, Bardy C, Reyna SM, Mu Y, Herrera C, et al. Probing sporadic and familial
657   Alzheimer's disease using induced pluripotent stem cells. Nature. 2012;482(7384):216-20.

658   8.   Ben-David U, Benvenisty N. The tumorigenicity of human embryonic and induced pluripotent
659   stem cells. Nat Rev Cancer. 2011;11(4):268-77.

660   9.   Kanemura H, Go MJ, Shikamura M, Nishishita N, Sakai N, Kamao H, et al. Tumorigenicity
661   studies of induced pluripotent stem cell (iPSC)-derived retinal pigment epithelium (RPE) for the
662   treatment of age-related macular degeneration. PLoS One. 2014;9(1):e85336.

663   10.   Okita K, Ichisaka T, Yamanaka S. Generation of germline-competent induced pluripotent stem
664   cells. Nature. 2007;448(7151):313-7.

665   11.   Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, et al. Role of the murine
666   reprogramming factors in the induction of pluripotency. Cell. 2009;136(2):364-77.

667   12.   Shu J, Wu C, Wu Y, Li Z, Shao S, Zhao W, et al. Induction of pluripotency in mouse somatic
668   cells with lineage specifiers. Cell. 2013;153(5):963-75.

669   13.   Cacchiarelli D, Trapnell C, Ziller MJ, Soumillon M, Cesana M, Karnik R, et al. Integrative
670   Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency. Cell.
671   2015;162(2):412-24.

672   14.   He X, Cao Y, Wang L, Han Y, Zhong X, Zhou G, et al. Human fibroblast reprogramming to
673   pluripotent stem cells regulated by the miR19a/b-PTEN axis. PLoS One. 2014;9(4):e95213.

674   15.   Huh S, Song HR, Jeong GR, Jang H, Seo NH, Lee JH, et al. Suppression of the ERK-SRF axis
675   facilitates somatic cell reprogramming. Exp Mol Med. 2018;50(2):e448.

676   16.   Miles DC, de Vries NA, Gisler S, Lieftink C, Akhtar W, Gogola E, et al. TRIM28 is an
677   Epigenetic Barrier to Induced Pluripotent Stem Cell Reprogramming. Stem Cells. 2017;35(1):147-57.

678   17.   Polo JM, Anderssen E, Walsh RM, Schwarz BA, Nefzger CM, Lim SM, et al. A molecular
679   roadmap of reprogramming somatic cells into iPS cells. Cell. 2012;151(7):1617-32.

680   18.   Dabiri Y, Gama-Brambila RA, Taskova K, Herold K, Reuter S, Adjaye J, et al. Imidazopyridines
681   as Potent KDM5 Demethylase Inhibitors Promoting Reprogramming Efficiency of Human iPSCs.
682   iScience. 2019;12:168-81.

30

683   19.   Hong H, Takahashi K, Ichisaka T, Aoi T, Kanagawa O, Nakagawa M, et al. Suppression of
684   induced pluripotent stem cell generation by the p53-p21 pathway. Nature. 2009;460(7259):1132-5.
685   20.   Robertson A, Mohamed TM, El Maadawi Z, Stafford N, Bui T, Lim DS, et al. Genetic ablation of
686   the mammalian sterile-20 like kinase 1 (Mst1) improves cell reprogramming efficiency and increases
687   induced pluripotent stem cell proliferation and survival. Stem Cell Res. 2017;20:42-9.
688   21.   Smith ZD, Nachman I, Regev A, Meissner A. Dynamic single-cell imaging of direct
689   reprogramming reveals an early specifying event. Nat Biotechnol. 2010;28(5):521-6.
690   22.   Zhang J, Nuebel E, Daley GQ, Koehler CM, Teitell MA. Metabolic regulation in pluripotent stem
691   cells during reprogramming and self-renewal. Cell Stem Cell. 2012;11(5):589-95.
692   23.   Li R, Liang J, Ni S, Zhou T, Qing X, Li H, et al. A mesenchymal-to-epithelial transition initiates
693   and is required for the nuclear reprogramming of mouse fibroblasts. Cell Stem Cell. 2010;7(1):51-63.
694   24.   Megyola CM, Gao Y, Teixeira AM, Cheng J, Heydari K, Cheng EC, et al. Dynamic migration
695   and cell-cell interactions of early reprogramming revealed by high-resolution time-lapse imaging. Stem
696   Cells. 2013;31(5):895-905.
697   25.   Dufour A, Thibeaux R, Labruyere E, Guillen N, Olivo-Marin JC. 3-D active meshes: fast discrete
698   deformable models for cell tracking in 3-D time-lapse microscopy. IEEE Trans Image Process.
699   2011;20(7):1925-37.
700   26.   Dzyubachyk O, van Cappellen WA, Essers J, Niessen WJ, Meijering E. Advanced level-set-based
701   cell tracking in time-lapse fluorescence microscopy. IEEE Trans Med Imaging. 2010;29(3):852-67.
702   27.   Maska M, Danek O, Garasa S, Rouzaut A, Munoz-Barrutia A, Ortiz-de-Solorzano C.
703   Segmentation and shape tracking of whole fluorescent cells based on the Chan-Vese model. IEEE
704   Trans Med Imaging. 2013;32(6):995-1006.
705   28.   Türetken E, Wang X, Becker CJ, Haubold C, Fua P. Network Flow Integer Programming to Track
706   Elliptical Cells in Time-Lapse Sequences. IEEE Transactions on Medical Imaging. 2017;36(4):942-51.
707   29.   Payer C, Štern D, Neff T, Bischof H, Urschler M, editors. Instance Segmentation and Tracking
708   with Cosine Embeddings and Recurrent Hourglass Networks. Medical Image Computing and
709   Computer Assisted Intervention – MICCAI 2018; 2018 2018//; Cham: Springer International
710   Publishing.
711   30.   Erdmann G, Volz C, Boutros M. Systematic approaches to dissect biological processes in stem
712   cells by image-based screening. Biotechnol J. 2012;7(6):768-78.
713   31.   Van Valen DA, Kudo T, Lane KM, Macklin DN, Quach NT, DeFelice MM, et al. Deep Learning
714   Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. PLoS
715   Comput Biol. 2016;12(11):e1005177.
716   32.   Chen CL, Mahjoubfar A, Tai LC, Blaby IK, Huang A, Niazi KR, et al. Deep Learning in
717   Label-free Cell Classification. Sci Rep. 2016;6:21471.
718   33.   Kraus OZ, Grys BT, Ba J, Chong Y, Frey BJ, Boone C, et al. Automated analysis of high-content
719   microscopy data with deep learning. Mol Syst Biol. 2017;13(4):924.
720   34.   Gao Z, Wang L, Zhou L, Zhang J. HEp-2 Cell Image Classification With Deep Convolutional
721   Neural Networks. IEEE J Biomed Health Inform. 2017;21(2):416-28.
722   35.   Stumpf PS, MacArthur BD. Machine Learning of Stem Cell Identities From Single-Cell
723   Expression Data via Regulatory Network Archetypes. Front Genet. 2019;10:2.
724   36.   Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System.   Proceedings of the 22nd
725   ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '162016.
726   p. 785-94.

727   37.   Chen J, Liu J, Chen Y, Yang J, Chen J, Liu H, et al. Rational optimization of reprogramming
728   culture conditions for the generation of induced pluripotent stem cells with ultra-high efficiency and
729   fast kinetics. Cell Res. 2011;21(6):884-94.
730   38.   Esteban MA, Wang T, Qin B, Yang J, Qin D, Cai J, et al. Vitamin C enhances the generation of
731   mouse and human induced pluripotent stem cells. Cell Stem Cell. 2010;6(1):71-9.
732   39.   Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al.
733   SCENIC: single-cell regulatory network inference and clustering. Nature Methods. 2017;14:1083.
734   40.   Li H, Pang F, Shi Y, Liu Z. Cell dynamic morphology classification using deep convolutional
735   neural networks. Cytometry A. 2018;93(6):628-38.
736   41.   Zhong J, Sun Y, Peng W, Xie M, Yang J, Tang X. XGBFEMF: An XGBoost-Based Framework
737   for Essential Protein Prediction. IEEE Trans Nanobioscience. 2018;17(3):243-50.
738   42.   Chen CLP, Zhang T, Chen L, Tam SC. I-Ching Divination Evolutionary Algorithm and its
739   Convergence Analysis. IEEE Transactions on Cybernetics. 2017;47(1):2-13.
740   43.   Zhang T, Chen CLP, Chen L, Xu X, Hu B. Design of Highly Nonlinear Substitution Boxes Based
741   on I-Ching Operators. IEEE Transactions on Cybernetics. 2018;48(12):3349-58.
742

Supporting information

S1 File. Supplementary Material

Fig1

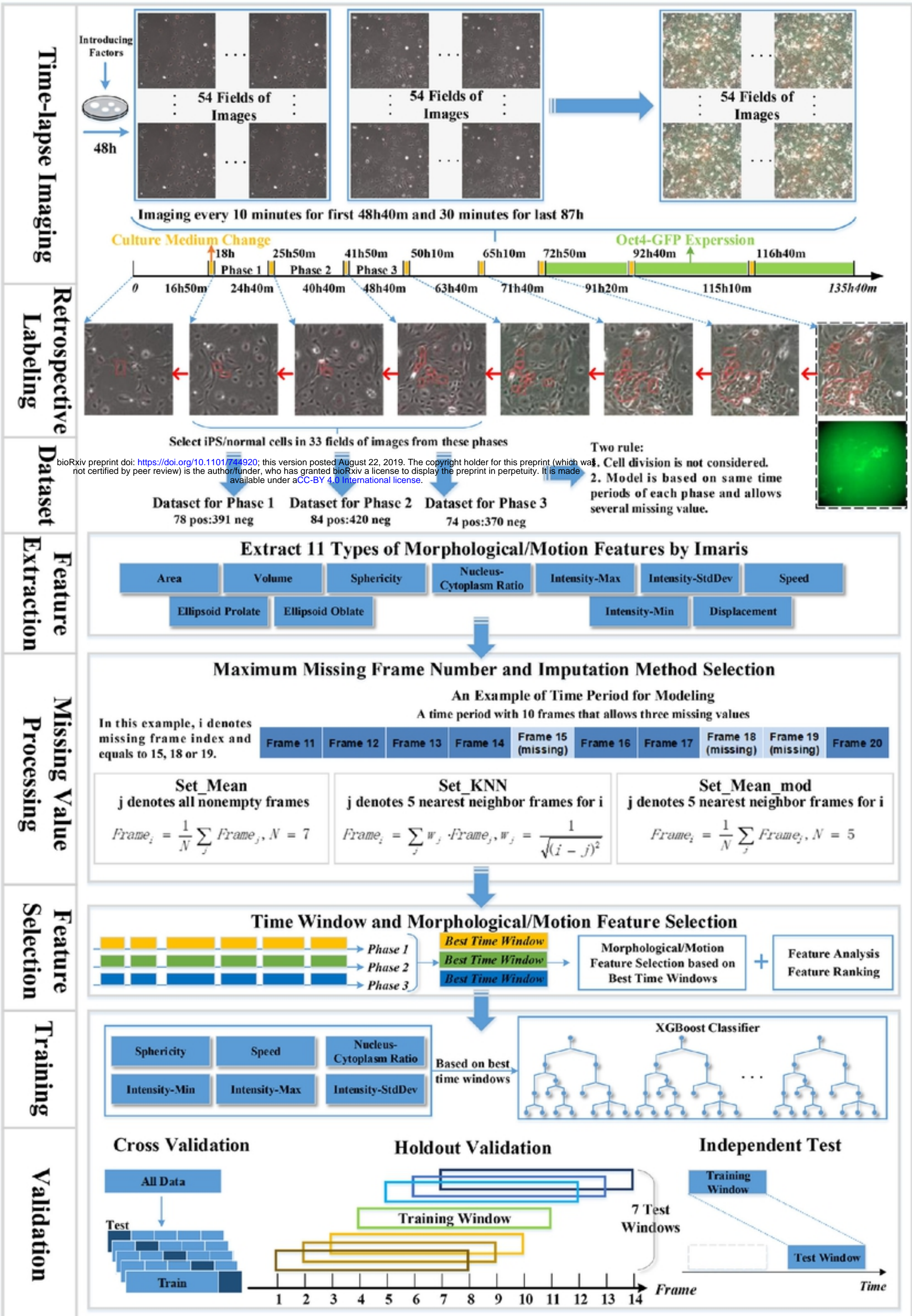Fig2

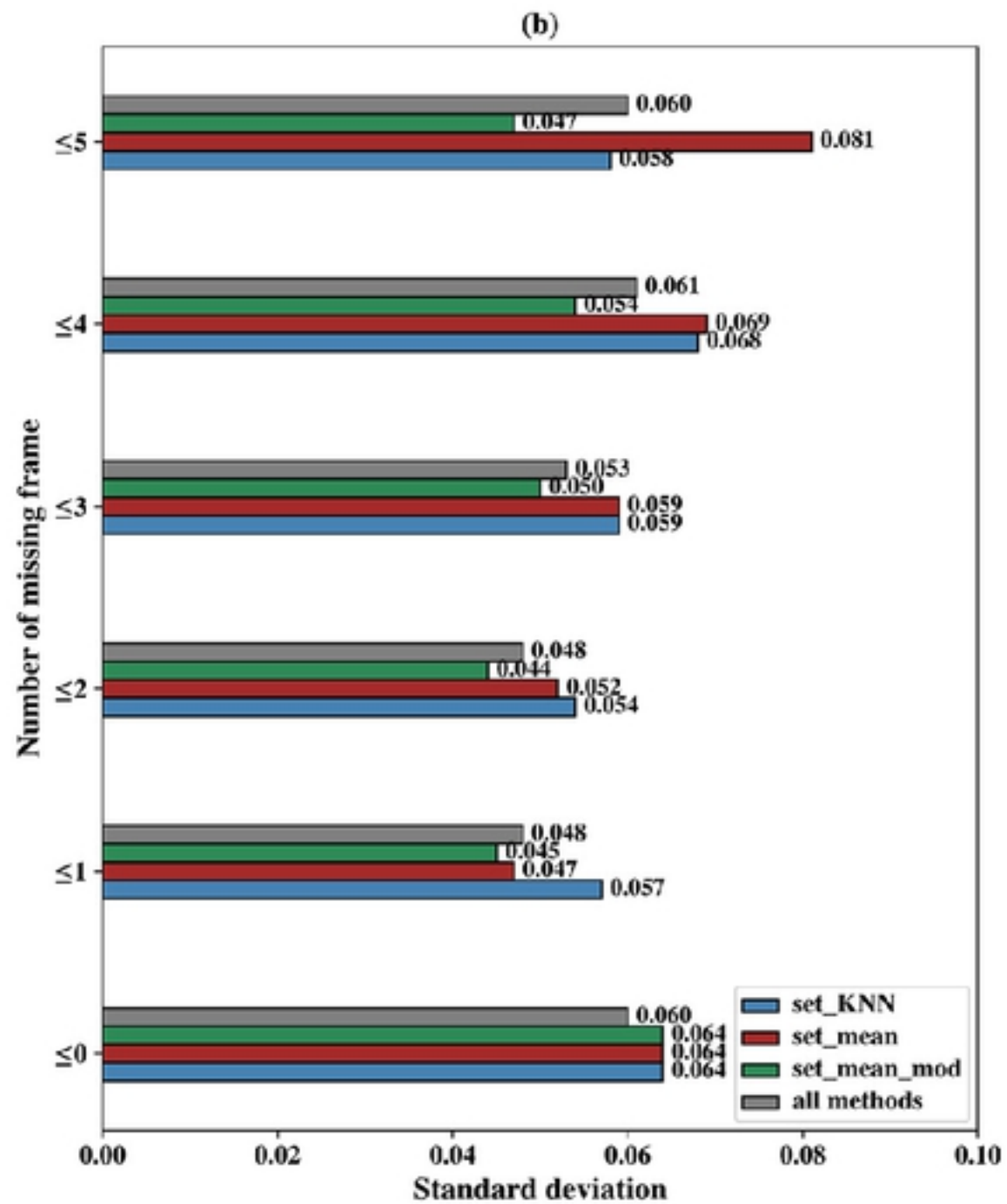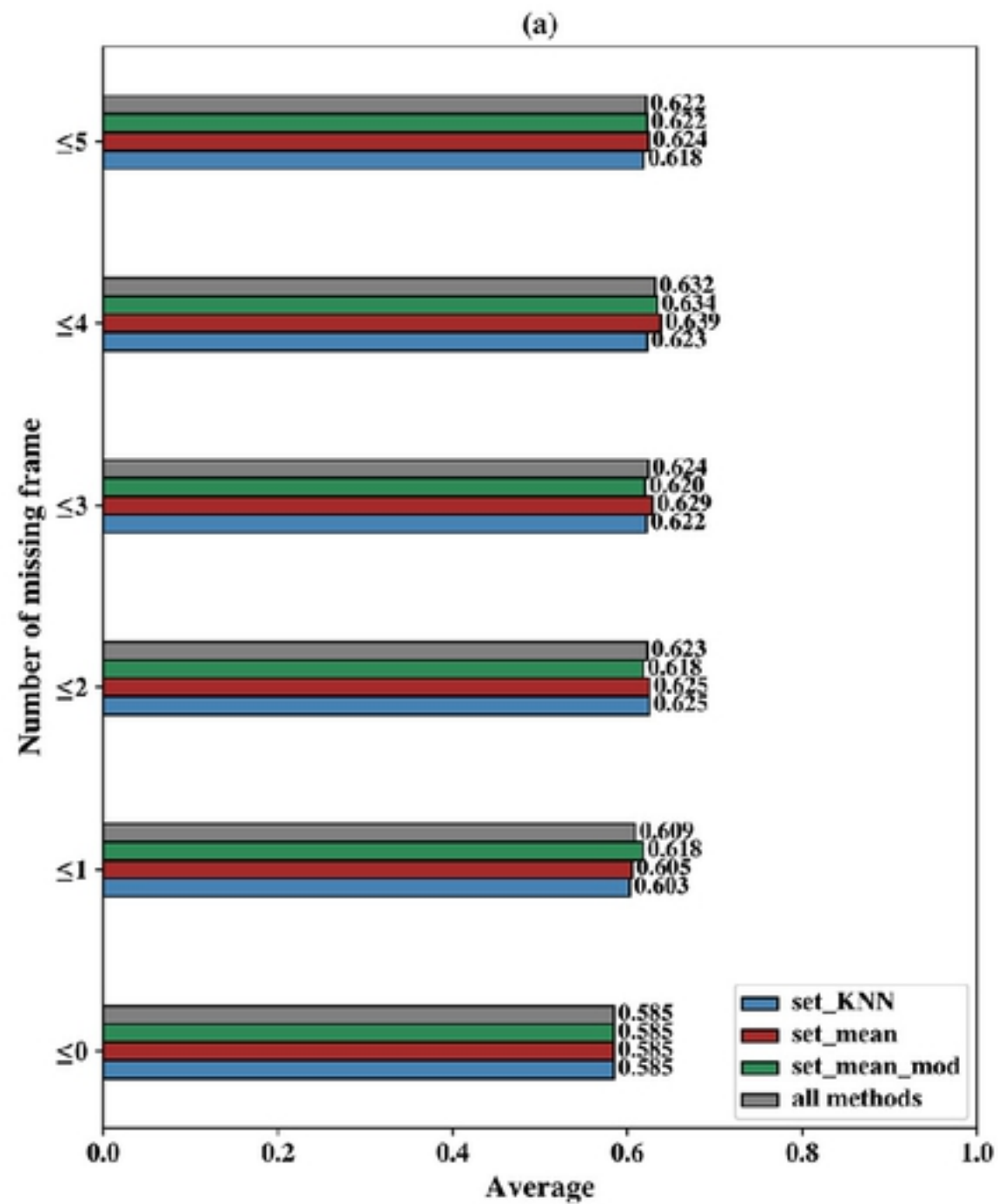Fig3
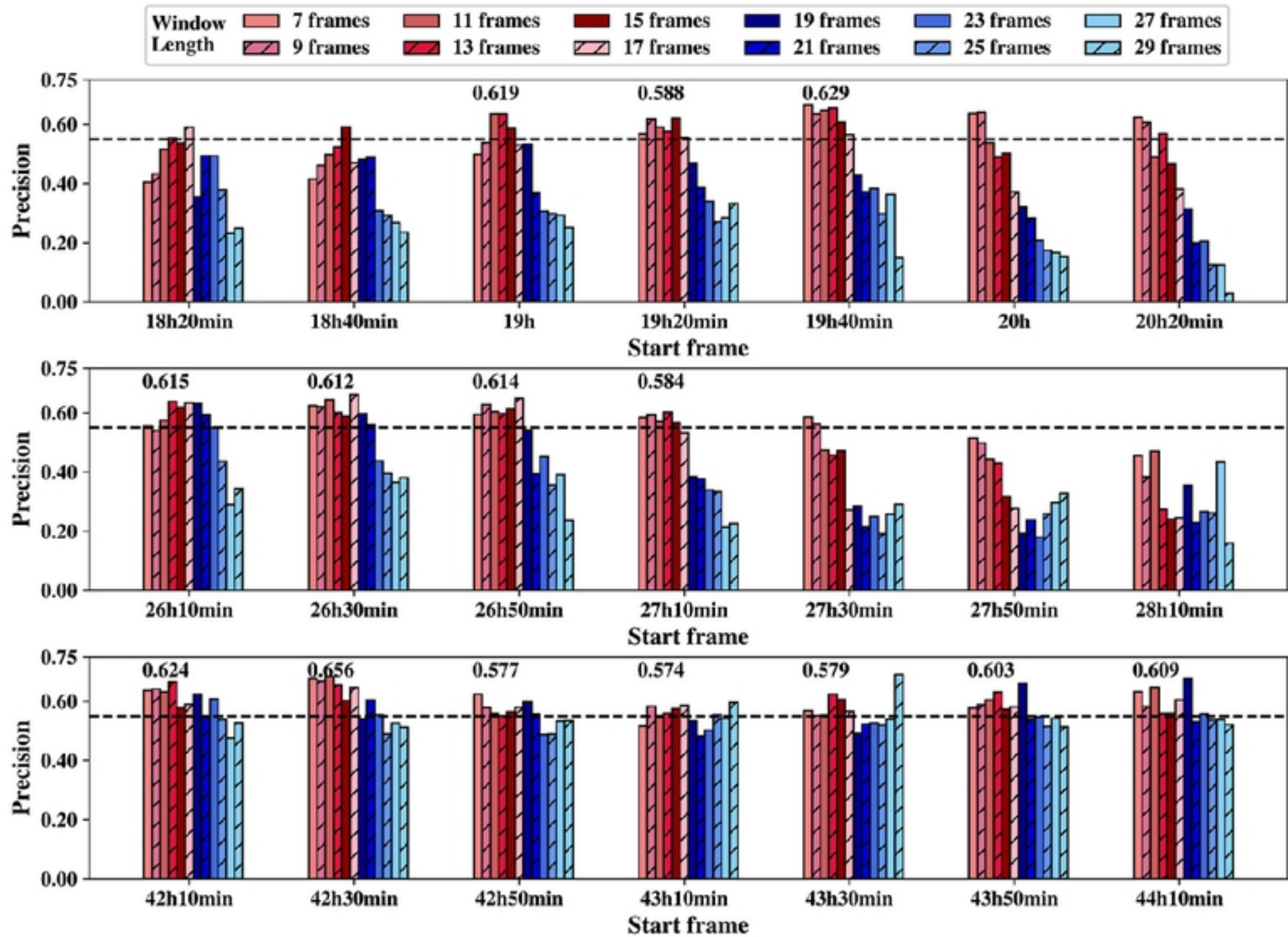
Fig4

(a)

(b)

Phase 1

	Sphericity	E-prolate	Area	Speed	Ratio	I-Max	I-stdDev	I-Min
Sphericity	1.00	0.66	0.77	0.10	0.18	0.51	0.49	0.38
E-prolate	0.66	1.00	0.32	0.03	0.17	0.24	0.25	0.29
Area	0.77	0.32	1.00	0.11	0.03	0.47	0.40	0.42
Speed	0.10	0.03	0.11	1.00	0.11	0.03	0.03	0.12
Ratio	0.18	0.17	0.03	0.11	1.00	0.30	0.37	0.28
I-Max	0.51	0.24	0.47	0.03	0.20	1.00	0.49	0.16
I-stdDev	0.49	0.25	0.40	0.03	0.27	0.49	1.00	0.09
I-Min	0.38	0.29	0.42	0.12	0.28	0.16	0.09	1.00

Phase 2

	Sphericity	E-prolate	E-oblate	Area	Speed	Ratio	I-Max	I-stdDev	I-Min
Sphericity	1.00	0.72	0.19	0.79	0.01	0.20	0.50	0.49	0.42
E-prolate	0.72	1.00	0.45	0.45	0.08	0.19	0.33	0.34	0.24
E-oblate	0.19	0.45	1.00	0.32	0.11	0.05	0.05	0.06	0.02
Area	0.79	0.45	0.32	1.00	0.01	0.08	0.45	0.38	0.36
Speed	0.01	0.08	0.11	0.01	1.00	0.03	0.02	0.01	0.03
Ratio	0.20	0.19	0.05	0.08	0.03	1.00	0.49	0.51	0.07
I-Max	0.50	0.33	0.05	0.45	0.02	0.49	1.00	0.46	0.08
I-stdDev	0.49	0.34	0.06	0.38	0.01	0.51	0.46	1.00	0.14
I-Min	0.42	0.24	0.02	0.36	0.03	0.07	0.08	0.14	1.00

Phase 3

	Sphericity	E-oblate	Speed	Ratio	I-Max	I-stdDev	I-Min
Sphericity	1.00	0.24	0.06	0.25	0.43	0.43	0.37
E-oblate	0.24	1.00	0.12	0.12	0.05	0.07	0.08
Speed	0.06	0.12	1.00	0.01	0.03	0.01	0.06
Ratio	0.25	0.12	0.01	1.00	0.55	0.57	0.09
I-Max	0.43	0.05	0.03	0.55	1.00	0.49	0.06
I-stdDev	0.43	0.07	0.01	0.57	0.49	1.00	0.16
I-Min	0.37	0.08	0.06	0.09	0.06	0.16	1.00

Pearson Coefficient

Fig5

Fig6