Journal of Pharmacological Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Pharmacological Sciences

journal homepage: www.elsevier.com/locate/jphs

Short Communication

Deep learning-based quality control of cultured human-induced pluripotent stem cell-derived cardiomyocytes

Ken Orita [a], Kohei Sawada [a], Ryuta Koyama [a], Yuji Ikegaya [a, b, *]

[a] Graduate School of Pharmaceutical Sciences, The University of Tokyo, Tokyo, 113-0033, Japan
[b] Center for Information and Neural Networks, National Institute of Information and Communications Technology, Suita City, Osaka, 565-0871, Japan

ARTICLE INFO

Article history:
Received 31 January 2019
Received in revised form
4 April 2019
Accepted 10 April 2019
Available online xxx

Keywords:
Machine learning
Heart
iPSC

ABSTRACT

Using bright-field images of cultured human-induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs), we trained a convolutional neural network (CNN), a machine learning technique, to decide whether the qualities of cell cultures are suitable for experiments. VGG16, an open-source CNN framework, resulted in a mean F1 score of 0.89 and judged the cell qualities at a speed of approximately 2000 images per second when run on a commercially available laptop computer equipped with Core i7. Thus, CNNs provide a useful platform for the high-throughput quality control of hiPSC-CMs.

Drug-induced cardiotoxicity, such as QT prolongation and torsades de pointes, is one of the main reasons for drug attrition[1] and contributes to 16% of the adverse effects reported in clinical trials.[2] Thus, establishing preclinical cardiac safety assay systems with high clinical predictability is of particular importance; however, due to species differences, conventional models for safety pharmacology studies using animal-derived tissues and cells have failed to fully predict drug responses in humans.[3] Human-induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs)[4] may help resolve this problem. However, one concern is the inconsistent qualities of hiPSC-CMs; the viability of cultured hiPSC-CMs varies depending on product lots, experimental trials, and the skills of experimenters, even though experimental procedures are identical, and some cultures fail to grow normally. To attain reproducibility, well-trained experimenters must inspect the nature of hiPSC-CMs visually and exclude low-quality hiPSC-CMs from data analyses with care. This manual process is restricted by human resources and prevents high-throughput screening for cardiotoxicity. Given that previous studies have shown that machine learning is useful for the quality control of hiPSC colonies,[5–7] we reasoned that a similar machine learning approach is also applicable to the quality control of hiPSC-CMs. To test this hypothesis, we utilized a convolutional neural network (CNN), which is a machine learning technique that was inspired by the visual system[8] and has been widely used for image classification. We trained a CNN model using bright-field images of two-dimensionally cultured hiPSC-CMs to accurately classify the images into 'normal' (i.e., useable in experiments) or 'abnormal' (unusable in experiments).

iCell Cardiomyocytes[2], or commercially available hiPSC-CMs, were obtained from Cellular Dynamics International (CDI; Madison, WI, USA). hiPSC-CMs were thawed according to the CDI protocol, suspended at a density of 400 cells/μL in the iCell-Cardiomyocytes plating medium (CDI), and seeded at 40,000 cells/well into fibronectin-coated 96-well plates (Corning, Corning, NY, USA). The plates were incubated at 37 °C under a 5% CO_2 atmosphere. The culture medium was refreshed after 24 h and, thereafter, half of the medium was refreshed every 2–3 d.

After 5–7 d of incubation, the cells were treated with doxorubicin (0.1, 0.3, and 1.0 μM; FUJIFILM Wako Pure Chemical Corporation, Osaka, Japan), sunitinib (0.1, 0.3, and 1.0 μM; Santa Crus Biotechnology), imatinib (0.3, 1.0, and 3.0 μM; Sigma–Aldrich, St. Louis, MO, USA), vandetanib (0.1, 0.3, and 1.0 μM; Toronto Research Chemicals, Toronto, Canada), nilotinib (0.1, 0.3, and 1.0 μM; Chemscene LLC, Monmouth Junction, NJ, USA) acetylsalicylic acid (10 μM; Nacalai Tesque, Kyoto, Japan), famotidine (10 μM; Sigma–Aldrich), or vehicle (0.1% DMSO; Sigma–Aldrich) solutions. These drug treatments were conducted to expand the variability in the qualities of hiPSC-CM cultures through their potential toxicity,

* Corresponding author. Laboratory of Chemical Pharmacology, Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan. Fax: +81 3 5841 4786.
E-mail address: yuji@ikegaya.jp (Y. Ikegaya).
Peer review under responsibility of Japanese Pharmacological Society.

https://doi.org/10.1016/j.jphs.2019.04.008

which may strengthen the robustness and versatility of machine learning. Bright-field images of cultured hiPSC-CMs (1280 × 1080 pixels, 16-bit intensity) were obtained before and 1–3 d after the drug treatments using a 20 × objective and CQ1 confocal quantitative image cytometer (Yokogawa Electric, Tokyo, Japan).

A total of 624 images were experimentally collected (Fig. 1A) and the qualities of the cultures were inspected by a well-trained experimenter and labeled as 'normal' (n = 556 images) or 'abnormal' (n = 68 images). The dataset was randomly split into 9 groups, and each group was used once as a testing dataset. The remaining 8 groups were split into training dataset (seven groups) and validation dataset (one group). The images in training, validation, and testing dataset were digitally increased to 14,000, 2000 and 2000 images, in which normal and abnormal images equally existed, respectively using data augmentation, in which randomly cropped 224 × 224-pixel sub-images were rotated 90°, 180°, 270°, or 360° and flipped vertically or horizontally (Fig. 1B). A total of grayscale 18,000 images were converted into RGB-colored photos using the Python library Pillow.

We used Chainer 4.3.0 to construct a VGG16 architecture (Fig. 2A) that had been pretrained using the ImageNet dataset[9] for transfer learning, which is a machine-learning technique that embodies effective learning when the number of datasets is limited.[10] The VGG16 architecture contains a total of 16 layers, which consist of 13 convolutional layers and 3 fully connected layers. The convolutional layers are used for extracting image features, such as edges, colors (low-level features), and faces (high-level features). The fully connected layers are used for image classification through nonlinear combination of the extracted image features. The max pooling layers perform down-sampling of image features to reduce the computational cost and obtain the robustness to translation invariant. The term "VGG" is derived from the name of developers' research group "Visual Geometry Group". All convolutional layers of VGG16 were frozen and used as fixed feature extractors, except for the eighth fully connected layer (fc8 layer). The number of nodes in the fc8 layer was changed from original 1000 nodes to 2 nodes for binary classification. To facilitate the calculation, the outputs (4096-dimensional feature vectors) of the fc7 layer were calculated for all 18,000 images and stored in advance. These fc7 feature vectors were visualized using uniform manifold approximation and projection for dimensional reduction (UMAP), which is a nonlinear dimensionality reduction algorithm based on a combination of Riemannian geometry and algebraic topology in the metric realization of fuzzy simplicial sets.[11,12] A total of 4096 dimensions of the feature vectors were reduced into two dimensions using UMAP (Python implemented with default parameters *n_neighbors* = 15, *min_dist* = 0.1, and *metric* = 'euclidean'). The data points of two classes were likely separated in the UMAP space (Fig. 2B; n = 18,000 images). This separation encouraged us to train the VGG16 model.

Using the training dataset, we trained the VGG16 fc8 layer on a CentOS7 computer equipped with an AMD Ryzen™ 5 1600 central processing unit, an NVIDIA GeForce GTX 1070 graphics processing

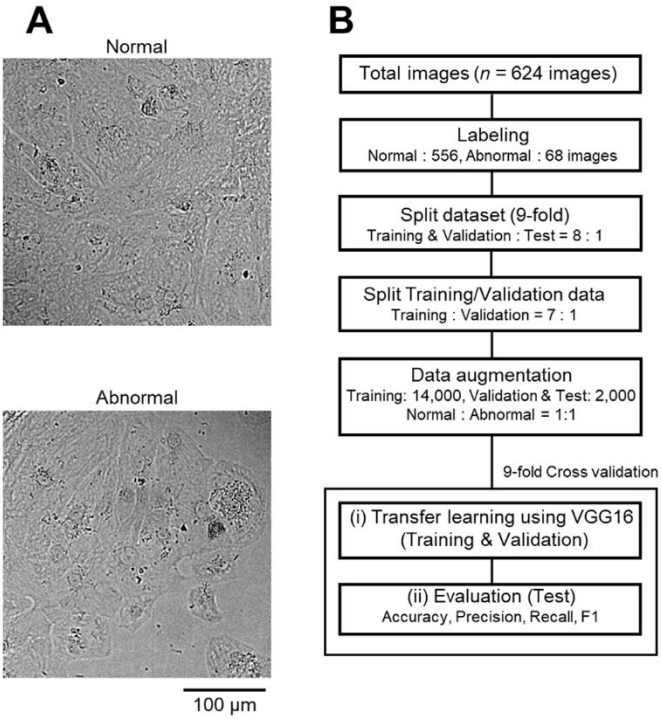

Fig. 1. Representative images of hiPSC-CMs and a workflow diagram for our experimental procedures. A, Bright-field images of hiPSC-CMs are labeled as 'normal' or 'abnormal'. B, Workflow diagram. Images were labeled and split into training (77.8%), validation (11.1%) and testing datasets (11.1%). Then images were data augmented. (i) VGG16 was retrained using the training sets, while its performance was tested using the validation sets. (ii) The post-training model was evaluated based on accuracy, precision, recall and F1 score using the testing set. To avoid accidental dependence on randomly assigned datasets, processes (i)-(ii) were repeated nine times, and the results are shown as the mean ± s.e.m. or the boxplots of the 9 models.

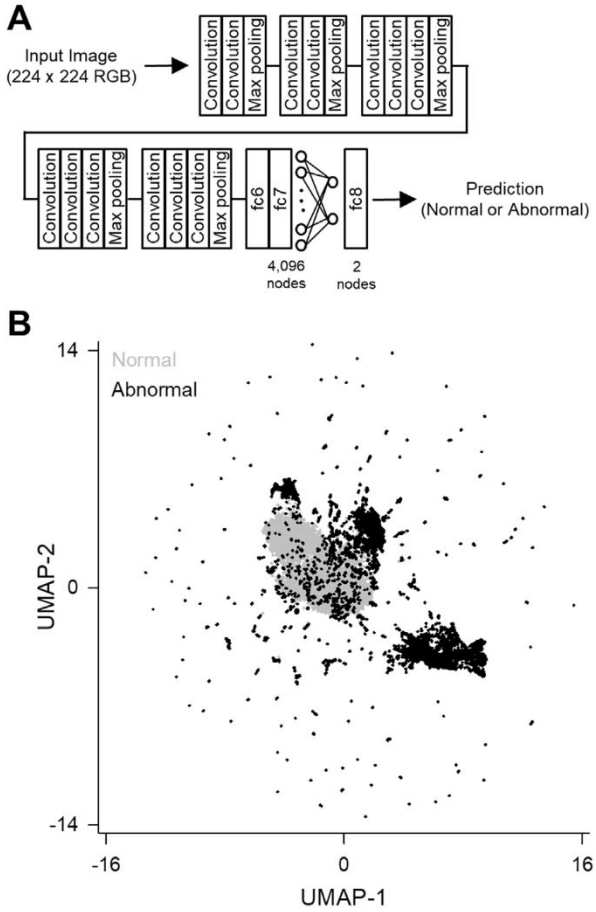

Fig. 2. The network architecture and UMAP visualization of the output of the VGG16 fc7 layer. A, The pretrained VGG16 model was used for transfer learning. Except for eighth fully connected layer (fc8 layer), all layers were frozen and used as a feature extractor. The number of nodes in the fc8 layer was reduced from 1000 to 2 and used for the binary prediction of normal or abnormal images. B, Uniform manifold approximation and projection (UMAP) visualization of the output of the VGG16 seventh fully connected layer (fc7 layer) for 9000 normal (gray) and 9000 abnormal images (black).

Please cite this article as: Orita K et al., Deep learning-based quality control of cultured human-induced pluripotent stem cell-derived cardiomyocytes, Journal of Pharmacological Sciences, https://doi.org/10.1016/j.jphs.2019.04.008

unit, and a 32 GB random access memory (RAM) to accurately predict two classes of images: normal or abnormal. The weights between the 4096-dimensional nodes of the fc7 layer and the two final nodes of fc8 layer were gradually updated using a stochastic gradient descent algorithm in batches of 2048 images per iteration via an AMSGrad Optimizer,[13] with a learning rate of 0.001. Learning was monitored based on the performance of the prediction with the validation datasets.

We evaluated our VGG16 model based on four parameters (accuracy, precision, recall, and F1 score), which are commonly used in the field of machine learning. Precision and recall reflect how false positive rates and false negative rates are small, respectively. Accuracy reflects how the overall predictions are correct, and F1, which is defined by the harmonic average of precision and recall and means the balance between precision and recall, is similar but usually more robust than accuracy when the datasets are imbalanced between classes. The accuracy increased gradually during training (Fig. 3A). To avoid overfitting, training was stopped at a maximum of 500 epochs unless the cross-entropy loss on the validation set decreased during the 20 successive epochs (Fig. 3B). One training session spent less than 5 min per 14,000 images. The parameters in the epoch where the cross-entropy loss on the validation set was the lowest were adopted as the final model.

To assess the performance of the post-training VGG16, we performed a 9-fold cross-validation using the testing dataset. On average, the accuracy, precision, recall, and F1 score were 0.897 ± 0.01, 0.946 ± 0.005, 0.843 ± 0.02, and 0.890 ± 0.01, respectively (Fig. 3C; mean ± 10,000 times bootstrapped s.e.m. of the 9 trained models). We also trained the same VGG16 model using randomly label-assigned images to estimate the level of prediction. The accuracy, precision, recall, and F1 score changed to 0.546 ± 0.02, 0.545 ± 0.02, 0.484 ± 0.05, and 0.507 ± 0.03, respectively (Fig. 3C; $n = 9$ models), where accuracy, precision, recall and the F1 score were significantly lower than the true parameters of the VGG16 that was trained using real datasets ($^*P < 0.001$; 10,000 permutation tests). Thus, our VGG16 model correctly predicted two classes: normal and abnormal. The calculation for the prediction used less than 0.1 s per 2000 images. Using a commercially available low-spec laptop computer equipped with Core i7-6700HQ and 4 GB RAM, the calculation speed was approximately 1 s per 2000 images.

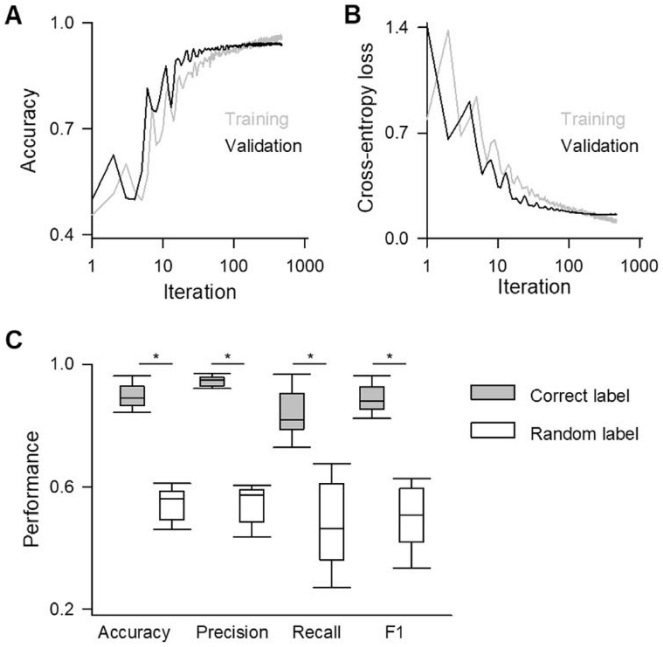

Fig. 3. Performances of the training, validation and testing datasets. A, B, Representative model performances during training. The accuracy (A) and cross-entropy loss (B) are plotted against the training iterations. The gray and black lines indicate the training datasets and validation datasets, respectively. C, Comparisons of accuracy, precision, recall, and F1 score of the 9-fold model cross-validation using testing dataset between the correct labels (gray) and random labels (white). $^*P < 0.001$ (10,000 permutation tests).

In this study, we established a method for the automated quality control of hiPSC-CMs. Certain explicit or implicit features differ between images of hiPSC-CMs with normal and abnormal qualities. Well experienced researchers can perceive such differences, but it is usually difficult or even impossible to orally communicate such discrimination skills (or intuitions). We discovered that after training, a CNN was able to successfully discriminate abnormal hiPSC-CMs from normal ones. Compared to humans, CNNs are dominant in terms of the cross-trial stability of task performance (reproducibility), speed of prediction, mental health in terms of work ethic, and unit labor cost. Indeed, our model performed at a speed of 2000 images per second, even on a low-cost computer purchased in a mass market.

Some limitations exist in this study. First, we used a single photography device. For universal usability of our approach, bright-field images taken by other devices will be needed for training and testing. Second, we trained only the VGG16 fc8 layer for transfer learning. Although there is a trade-off between the calculation speed and performance, full retraining of VGG16 will be effective for a better performance. A previous study demonstrated that the performances of CNNs depend on pretrained models.[10] More advanced pretrained models may enhance the overall performance. Third, we used random data augmentation to increase the number of images. During this process, similar images might have been generated and made the classification easier. More images will be needed to test model robustness.

In summary, we demonstrated, for the first time, that CNNs can assess the qualities of cultured hiPSC-CMs; otherwise, only well-experienced experts can assess these qualities. We hope that automated quality control systems using machine learning work not only as a decision supporting tool for humans but also as a complete alternative to humans. Our work is the first step toward this end.

Conflict of interest

The authors declare no conflicts of interest associated with this manuscript.

Acknowledgments

This work was supported by JST ERATO (JPMJER1801), JSPS Grants-in-Aid for Scientific Research (18H05525), and the Human Frontier Science Program (RGP0019/2016). This work was conducted as part of a program at the International Research Center for Neurointelligence (WPI-IRCN) at the University of Tokyo Institutes for Advanced Study at the University of Tokyo.

References

1. Stevens JL, Baker TK. The future of drug safety testing: expanding the view and narrowing the focus. *Drug Discov Today*. 2009;14:162–167.
2. Olson H, Betton G, Robinson D, et al. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol*. 2000;32:56–67.
3. Gintant G, Sager PT, Stockbridge N. Evolution of strategies to improve preclinical cardiac safety testing. *Nat Rev Drug Discov*. 2016;15:457–471.
4. Zhang J, Wilson GF, Soerens AG, et al. Functional cardiomyocytes derived from human induced pluripotent stem cells. *Circ Res*. 2009;104:e30–e41.

5. Joutsijoki H, Haponen M, Rasku J, Aalto-Setala K, Juhola M. Machine learning approach to automated quality identification of human induced pluripotent stem cell colony images. *Comput Math Methods Med*. 2016;2016:3091039.
6. Kavitha MS, Kurita T. Ahn BC Critical texture pattern feature assessment for characterizing colonies of induced pluripotent stem cells through machine learning techniques. *Comput Biol Med*. 2018;94:55–64.
7. Kavitha MS, Kurita T, Park SY, Chien SI, Bae JS. Ahn BC Deep vector-based convolutional neural network approach for automatic recognition of colonies of induced pluripotent stem cells. *PLoS One*. 2017;12, e0189974.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
9. Simonyan K, Zisserman A. *Very deep convolutional networks for large-scale image recognition*. 2014. arXiv preprint arXiv:1409.1556.
10. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172:1122–1131. e1129.
11. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2019;37:38–44.
12. McInnes L, Healy J, Melville J. *UMAP: uniform manifold approximation and projection for dimension reduction*. 2018. arXiv preprint arXiv:1802.03426.
13. Reddi SJ, Kale S, Kumar S. On the convergence of adam and beyond. In: *International Conference on Learning Representations*; 2018. https://openreview.net/pdf?id=ryQu7f-RZ.

Please cite this article as: Orita K et al., Deep learning-based quality control of cultured human-induced pluripotent stem cell-derived cardiomyocytes, Journal of Pharmacological Sciences, https://doi.org/10.1016/j.jphs.2019.04.008