Stem Cell Reports
Report

OPEN ACCESS

Automated Deep Learning-Based System to Identify Endothelial Cells Derived from Induced Pluripotent Stem Cells

Dai Kusumoto,[1,2,4] Mark Lachmann,[1,4] Takeshi Kunihiro,[3,4] Shinsuke Yuasa,[1,*] Yoshikazu Kishino,[1] Mai Kimura,[1] Toshiomi Katsuki,[1] Shogo Itoh,[1] Tomohisa Seki,[1,2] and Keiichi Fukuda[1]

[1]Department of Cardiology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan
[2]Department of Emergency and Critical Care Medicine, Keio University School of Medicine, Tokyo 160-8582, Japan
[3]LE Development Department, R&D Division, Medical Business Group, Sony Imaging Products & Solutions Inc., 4-14-1 Asahi-cho, Atsugi-shi, Kanagawa 243-0014, Japan
[4]Co-first author
*Correspondence: yuasa@keio.jp
https://doi.org/10.1016/j.stemcr.2018.04.007

SUMMARY

Deep learning technology is rapidly advancing and is now used to solve complex problems. Here, we used deep learning in convolutional neural networks to establish an automated method to identify endothelial cells derived from induced pluripotent stem cells (iPSCs), without the need for immunostaining or lineage tracing. Networks were trained to predict whether phase-contrast images contain endothelial cells based on morphology only. Predictions were validated by comparison to immunofluorescence staining for CD31, a marker of endothelial cells. Method parameters were then automatically and iteratively optimized to increase prediction accuracy. We found that prediction accuracy was correlated with network depth and pixel size of images to be analyzed. Finally, K-fold cross-validation confirmed that optimized convolutional neural networks can identify endothelial cells with high performance, based only on morphology.

INTRODUCTION

Machine learning consists of automated algorithms that enable learning from large datasets to resolve complex problems, including those encountered in medical science (Gorodeski et al., 2011; Heylman et al., 2015; Hsich et al., 2011). In deep learning, a form of machine learning, patterns from several types of data are automatically extracted (Lecun et al., 2015) to accomplish complex tasks such as image classification, which in conventional machine learning requires feature extraction by a human expert. Deep learning eliminates this requirement by identifying the most informative features using multiple layers in neural networks, i.e., deep neural networks (Hatipoglu and Bilgin, 2014), which were first conceived in the 1940s to mimic human neural circuits (McCulloch and Pitts, 1943). In such neural networks, each neuron receives weighted data from upstream neurons, which are then processed and transmitted to downstream neurons. Ultimately, terminal neurons calculate a predicted value based on processed data, and weights are then iteratively optimized to increase the agreement between predicted and observed values. This technique is rapidly advancing due to innovative algorithms and improved computing power (Bengio et al., 2006; Hinton et al., 2006). For example, convolutional neural networks have now achieved almost the same accuracy as a clinical specialist in diagnosing diabetic retinopathy and skin cancer (Esteva et al., 2017; Gulshan et al., 2016). Convolutional neural networks have also proved useful in cell biology such as morphological classification of hematopoietic cells, C2C12 myoblasts, and induced pluripotent stem cells (iPSCs) (Buggenthin et al., 2017; Niioka et al., 2018; Yuan-Hsiang et al., 2017).

iPSCs, which can be established from somatic cells by expression of defined genes (Takahashi and Yamanaka, 2006), hold great promise in regenerative medicine (Yuasa and Fukuda, 2008), disease modeling (Tanaka et al., 2014), drug screening (Avior et al., 2016), and precision medicine (Chen et al., 2016). iPSCs can differentiate into numerous cell types, although differentiation efficiencies vary among cell lines and are sensitive to experimental conditions (Hu et al., 2010; Osafune et al., 2008). In addition, differentiated cell types are difficult to identify without molecular techniques such as immunostaining and lineage tracing. We hypothesized that phase-contrast images contain discriminative morphological information that can be used by a convolutional neural network to identify endothelial cells. Accordingly, we investigated whether deep learning techniques can be used to identify iPSC-derived endothelial cells automatically based only on morphology.

RESULTS

Development of an Automated System to Identify Endothelial Cells

We differentiated iPSCs as previously described (Patsch et al., 2015), obtaining mesodermal cells at around day 3 and specialized endothelial cells at around day 5 (Figure S1A). At day 6, structures that resemble vascular tubes were formed (Figure S1B). CD31 staining confirmed that endothelial cells were obtained at an efficiency of 20%–35%, as assessed by flow cytometry. Differentiation efficiency was strongly variable (Figure S1C), highlighting the need for an automated cell identification system to assess iPSC differentiation or to identify and quantify the cell types formed.

The basic strategy to identify endothelial cells by convolutional neural networks is shown in Figure 1A. In brief, differentiated iPSCs were imaged by phase contrast and by immunofluorescence staining for CD31, a marker of endothelial cells. The latter were then binarized into white and black pixels corresponding to raw pixels above and below a threshold value, respectively. Subsequently, input blocks were extracted randomly from phase-contrast images, and matching target blocks equivalent to or within input blocks were extracted from both phase-contrast and binarized immunofluorescence images. Binarized target blocks were then classified as unstained (0) or stained (1) depending on the ratio of white pixels to black, to generate answers. Finally, input blocks were analyzed in LeNet, a small network (Lecun et al., 1998), and AlexNet, a large network (Krizhevsky et al., 2012), to predict phase-contrast target blocks as unstained or stained. Predictions were compared with answers obtained from binarized target blocks, and weights were automatically and iteratively optimized to train the neural networks and thereby increase accuracy (Figure 1A).

Networks were then optimized according to Figure 1B. Number of blocks, input block size, and target block size were first optimized using the small network, along with staining threshold, the ratio of white pixels to black for a target block to be classified as stained. To improve performance, as assessed by F1 score and accuracy, the small network was compared with the large network, observed errors were analyzed, and binarized target blocks were rebinarized by visual comparison of raw fluorescent images with phase-contrast images. Finally, the optimized network was validated by K-fold cross-validation (Figure 1B). To this end, we obtained 200 images from each of four independent experiments, of which 640 were used for training and 160 for validation to collect data shown in Figures 2 and 3. From each image, 200 blocks were randomly extracted, and 500–128,000 of the blocks were used for training while 32,000 blocks were used for validation (Figure 1C).

Improvement of F1 Score and Accuracy by Optimization

To train the networks we optimized several experimental conditions, including number of input blocks, target block size, and input block size. Performance was evaluated based on F1 scores, which aggregates recall and precision, and on accuracy, which is the fraction of correct predictions. As noted, we first used 500–128,000 blocks for training (Figure 1C) to determine the number of blocks required to achieved convergence (Table S1). Inflection points in F1 scores and accuracy were observed at 16,000 blocks, and convergence was achieved at 32,000 blocks for an input and target block size of 128 × 128 pixels, as well as for an input block size of 512 × 512 pixels and a target block size of 32 × 32 pixels (Figure 2A). Hence, 32,000 blocks were used for training in subsequent experiments. Next, the optimal combination of block size and staining threshold was determined by input blocks of 32 × 32, 64 × 64, 128 × 128, 256 × 256, and 512 × 512 pixels. We note that 32 × 32-pixel blocks contained only single cells, while 512 × 512-pixel blocks contained entire colonies and surrounding areas (Figure S2A). Based on F1 scores, performance was best from an input block size of 512 × 512 pixels combined with a staining threshold of 0.3 (Figures 2B and 2C; Table S2). Both F1 score and accuracy increased with input block size (Figures 2D, S2B, and S2C), indicating that areas surrounding cells should be included to increase accuracy. In contrast, target block size did not affect predictive power (Figure 2E) or the correlation between input block size and F1 scores and accuracy (Figure S2D and Table S3).

Effect of Network Size on Predictive Power

As network architecture is critical to performance, we compared the predictive power of the small network LeNet (Lecun et al., 1998) after training on 128,000 blocks with that of the large network AlexNet (Krizhevsky et al., 2012) (Figure 3A). F1 scores and accuracy from the latter were higher (Figures 3B and S3A), suggesting that extraction of complex features by a large network improves cell identification by morphology. Performance was further enhanced by analyzing true positives, true negatives, false positives, and false negatives (Figures 3C and S3B). We found that true positives and true negatives were typically obtained in areas with uniformly distributed cells. In contrast, areas with heterogeneous appearance, such as at the border between abundantly and sparsely colonized surfaces, often led to false positives or false negatives. To examine whether F1 scores are influenced by heterogeneous appearance (Figure S4A), we scored the complexity of all 32,000 512 × 512-pixel validation blocks as the average difference between adjacent pixels, normalized to the dynamic range (Saha and Vemuri, 2000). Blocks with complexity of <0.04 were considered sparsely colonized, while blocks with complexity of 0.04 to 0.08 typically contained uniformly distributed cells with clear boundaries. All other images had complexity >0.08 and contained dense colonies with indistinct cell borders. In both the small and large networks (Figures S4B, S4C, and S4D), F1 scores were highest for blocks with complexity of 0.04 to 0.08 (typically 0.06), implying that variations in cell

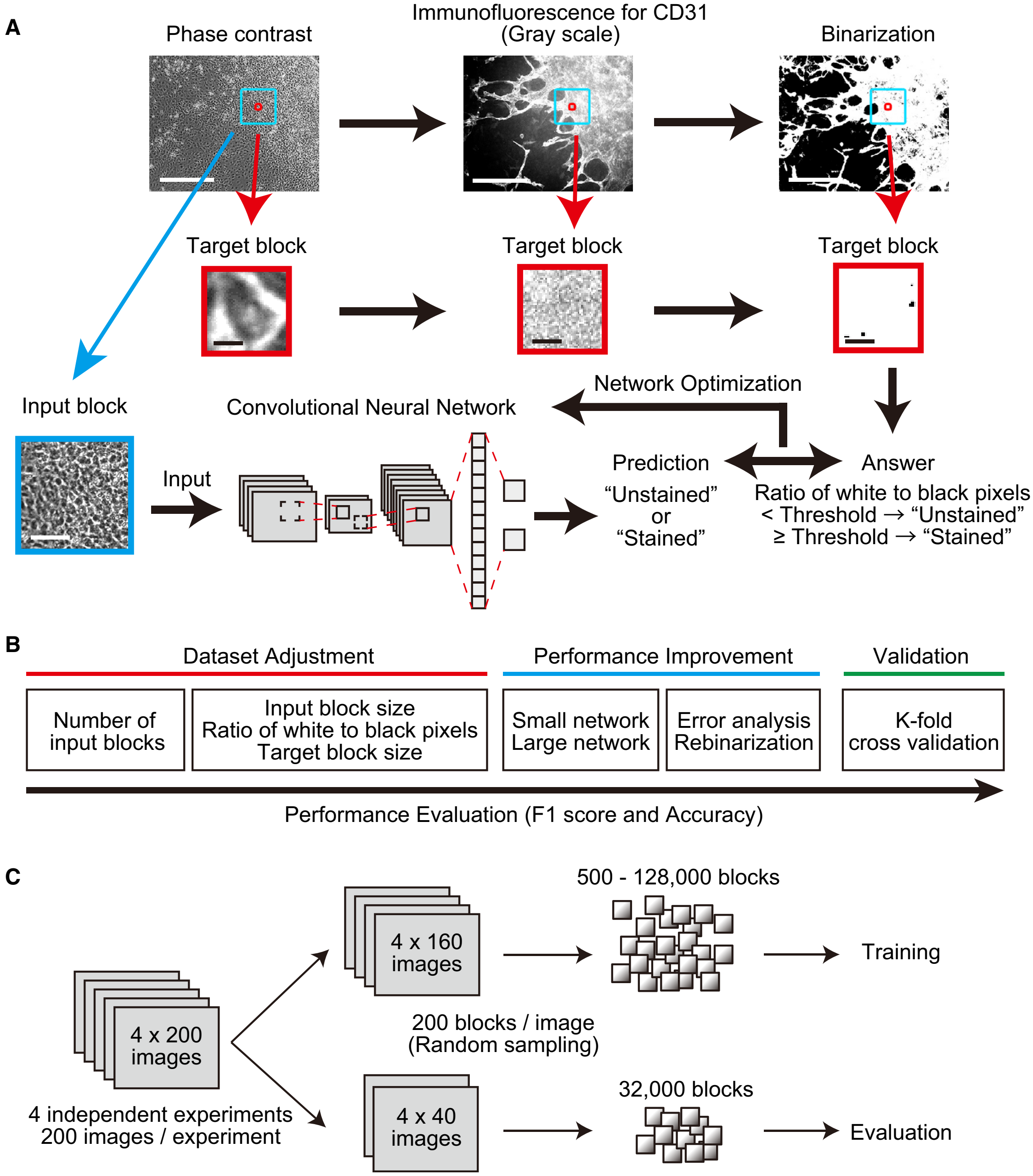
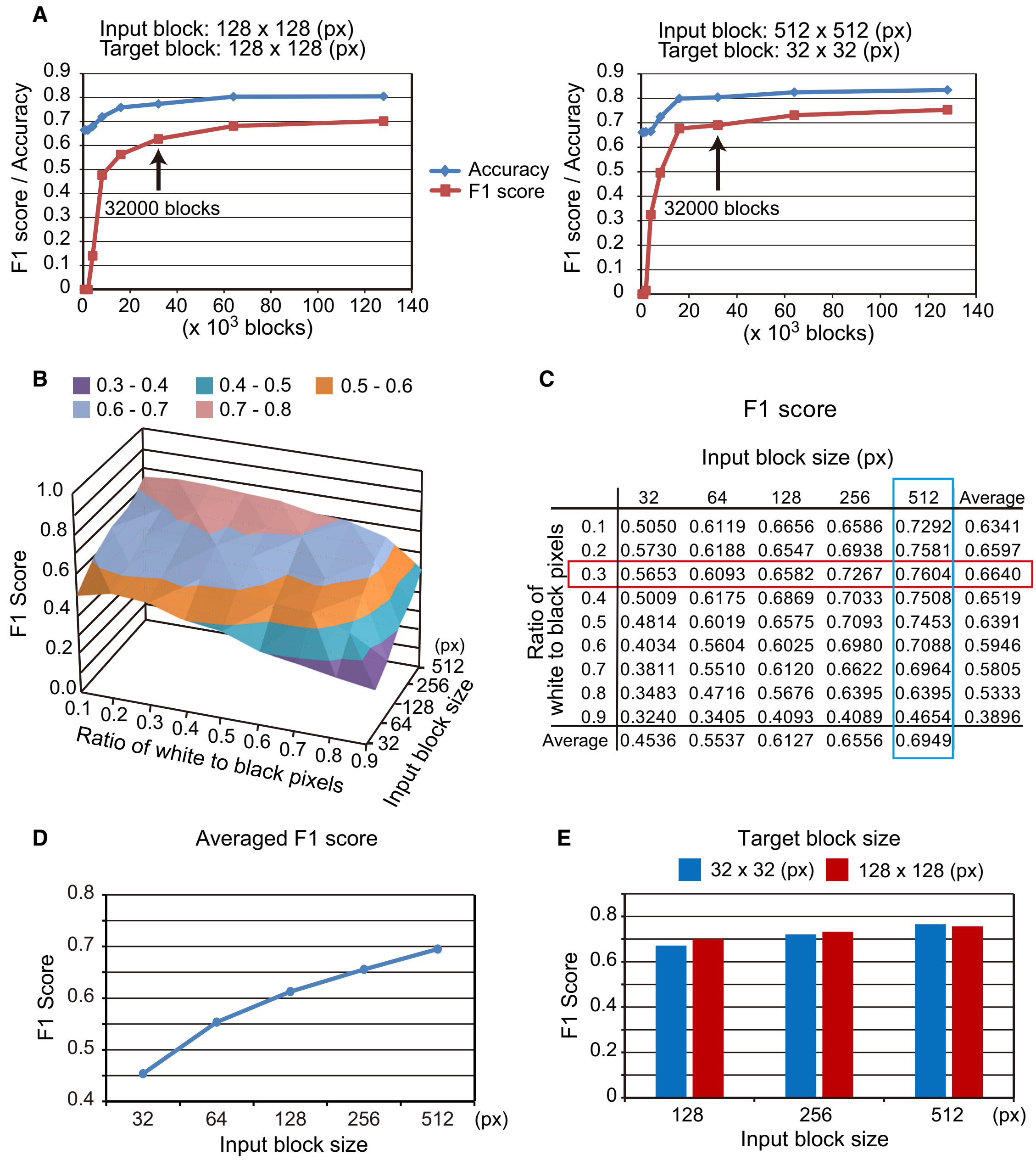


Figure 1. Analysis of Induced Pluripotent Stem Cell-Derived Endothelial Cells Using Convolutional Neural Networks
(A) Training protocol. Input blocks were extracted from phase-contrast images and predicted by networks to be unstained (0) or stained (1) for CD31. Target blocks containing single cells were extracted from immunofluorescent images of the same field, binarized based on CD31 staining, and classified as stained or unstained based on the ratio of white pixels to black. Network weights were then automatically and iteratively adjusted to maximize agreement between predicted and observed classification. Scale bars, 400 μm (upper panels), 5 μm (middle panels), and 80 μm (bottom panels).
(B) Optimization of experimental parameters to maximize F1 score and accuracy.
(C) Two hundred images each were obtained from four independent experiments. Images were randomized at 80:20 ratio into training and evaluation sets, and 200 blocks were randomly extracted from each image.



Ratio of white to black pixels	32	64	128	256	512	Average
0.1	0.5050	0.6119	0.6656	0.6586	0.7292	0.6341
0.2	0.5730	0.6188	0.6547	0.6938	0.7581	0.6597
0.3	0.5653	0.6093	0.6582	0.7267	0.7604	0.6640
0.4	0.5009	0.6175	0.6869	0.7033	0.7508	0.6519
0.5	0.4814	0.6019	0.6575	0.7093	0.7453	0.6391
0.6	0.4034	0.5604	0.6025	0.6980	0.7088	0.5946
0.7	0.3811	0.5510	0.6120	0.6622	0.6964	0.5805
0.8	0.3483	0.4716	0.5676	0.6395	0.6395	0.5333
0.9	0.3240	0.3405	0.4093	0.4089	0.4654	0.3896
Average	0.4536	0.5537	0.6127	0.6556	0.6949	

Figure 2. Dataset Adjustment

(A) F1 score and accuracy as a function of number of input blocks. Left: network performance using 128 × 128-pixel (px) input blocks and 128 × 128-px target blocks. Right: performance using 512 × 512-px input blocks and 32 × 32-px target blocks.

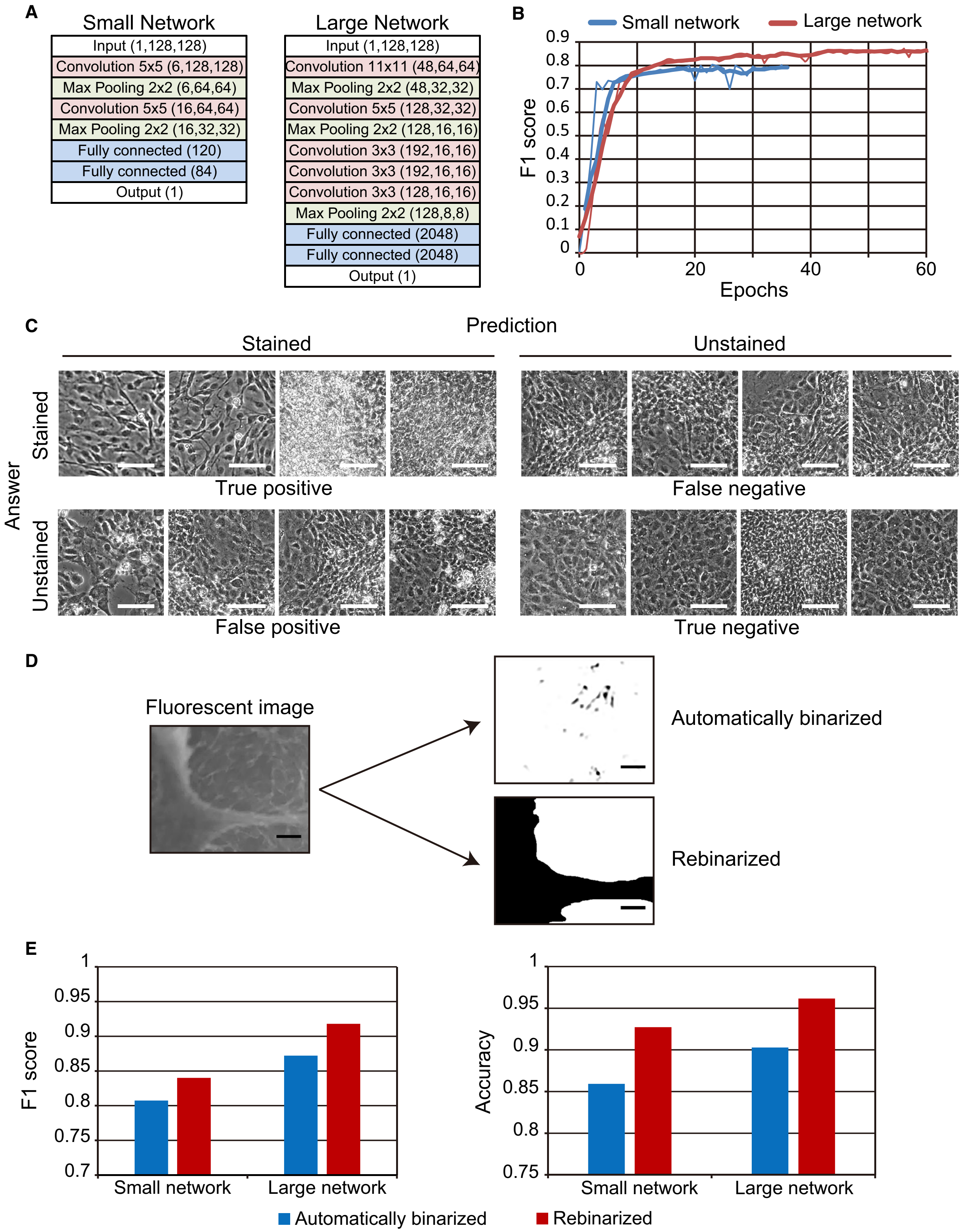
(B and C) F1 score as a function of input block size and staining threshold. The optimal threshold is boxed in red and the optimal input block size is boxed in blue.

(D) Average F1 score for different input block sizes.

(E) F1 score for different target block sizes.

See also Figure S2 and Tables S1–S3.

density and morphology affect network performance, in line with incorrect predictions as shown in Figures 3C and S3B. In light of this result, we speculated that weak staining, non-specific fluorescence, and autofluorescence in dense colonies may also degrade performance. Accordingly, we rebinarized target blocks by visual comparison



(legend on next page)

with raw fluorescent images (Figure 3D). Following this step, 26,861 of 128,000 blocks (21%) were classified as stained, while fully automated binarization scored 40,852 of 128,000 blocks (32%) as stained (Table S4A). Notably, the F1 score and accuracy rose above 0.9 and 0.95, respectively, in the large network (Figure 3E and Table S4A).

K-Fold Cross-Validation

Finally, we assessed network performance and generalization by K-fold cross-validation, in which k subsets of data are divided into $k - 1$ training datasets and one validation dataset. Training and validation are then performed k times using different combinations of training and validation datasets. In our case, 800 images were collected in four independent experiments, of which various combinations of 600 images from three experiments were used for training and 200 images from one experiment were used for validation (Figure 4A). The F1 score and accuracy were approximately 0.7 and higher than 0.7 for the small network with automatically binarized target blocks, but over 0.75 and over 0.9, respectively, for the large network with rebinarized target blocks (Figures 4B and 4C; Table S4B).

DISCUSSION

In this study, we demonstrated that deep learning techniques are effective in identifying iPSC-derived endothelial cells. Following optimization of parameters such as number of input blocks, target block size, input block size, staining threshold, and network size, we achieved satisfactory F1 scores and accuracy. Notably, we found that a larger input block increases prediction accuracy, indicating that the environment surrounding cells is an essential feature, as was also observed for differentiated C2C12 myoblasts (Niioka et al., 2018). We note that the immediate microenvironment is also an essential determinant of differentiation (Adams and Alitalo, 2007; Lindblom et al., 2003; Takakura et al., 2000), and that the positive correlation between input block size and F1 score or accuracy may prove helpful in future strategies to identify differentiated cells by morphology.

In comparison with other machine learning techniques, deep learning is straightforward and achieves high accuracies. Indeed, deep learning algorithms have won the ImageNet Large-Scale Visual Recognition Challenge since 2012 (He et al., 2015; Krizhevsky et al., 2012; Szegedy et al., 2014; Zeng et al., 2016), and have also proved useful in cell biology (Buggenthin et al., 2017; Niioka et al., 2018; Van Valen et al., 2016; Yuan-Hsiang et al., 2017). Although we used the older-generation networks LeNet and AlexNet, newer networks achieve even better accuracy in image classification (Esteva et al., 2017; Gulshan et al., 2016). Several techniques, such as increasing network depth (Simonyan and Zisserman, 2014), residual learning (He et al., 2015), and batch normalization (Ioffe and Szegedy, 2015), may also enhance performance, although these were not implemented in this study, since results were already satisfactory.

Inspection revealed some issues in binarizing heterogeneous areas in images with weak staining, non-specific fluorescence, and autofluorescence. To lower the number of false positives and improve performance, we rebinarized these images by comparing raw fluorescent images with phase-contrast images. In addition, cell density significantly affected F1 scores, implying that cells should be cultured carefully to a suitable density, or that networks should be trained to distinguish between true and false positives, especially when images are heterogeneous. Finally, K-fold cross-validation showed that iPSC-derived endothelial cells were identified with accuracy approximately 0.9 and F1 score 0.75, in line with similar attempts (Buggenthin et al., 2017; Niioka et al., 2018; Yuan-Hsiang et al., 2017).

Importantly, the data show that iPSC-derived endothelial cells can be identified based on morphology alone, requiring only 100 μs per block in a small network and 275 μs per block in a large network (Figure S4E). As morphology-based identification does not depend on labeling, genetic manipulation, or immunostaining, it can be used for various applications requiring native, living cells. Thus, this system may enable analysis of large datasets and advance cardiovascular research and medicine.

EXPERIMENTAL PROCEDURES

iPSC Culture

iPSCs were maintained in mTeSR with 0.5% penicillin/streptomycin on culture dishes coated with growth factor-reduced

Figure 3. Network Optimization

(A) Comparison of LeNet and AlexNet, which are small and large deep neural networks.

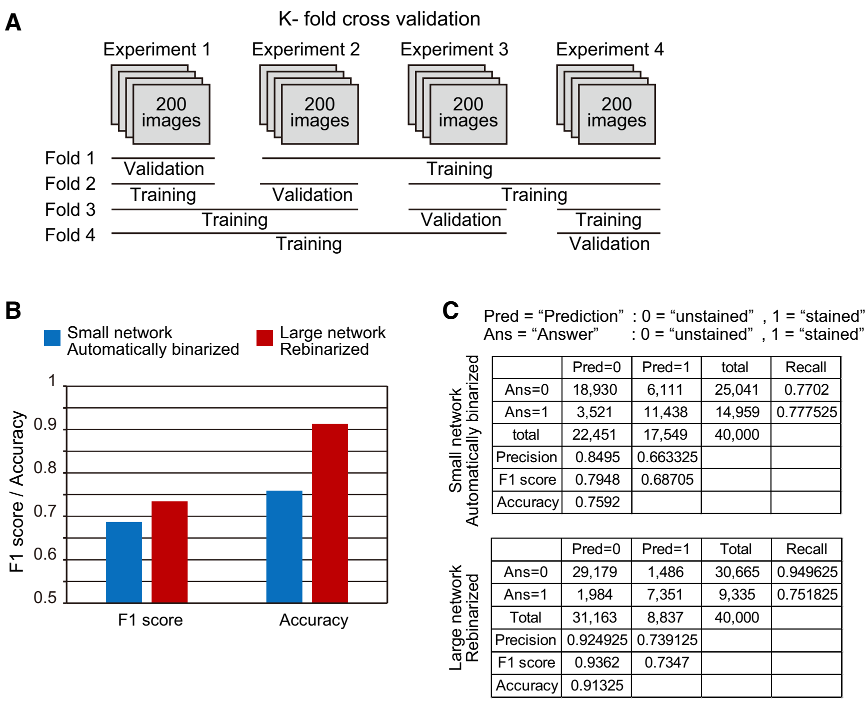
(B) F1 score learning curves from the small and large network.

(C) Representative true positive, false positive, true negative, and false negative images. Scale bars, 80 μm.

(D) Immunofluorescent images were binarized automatically, or rebinarized by manual comparison of raw fluorescent images to phase-contrast images. Scale bars, 100 μm.

(E) F1 score and accuracy were compared following training of the small and large network on automatically binarized or rebinarized target blocks.

See also Figures S3 and S4; Table S4.



Small network Automatically binarized

	Pred=0	Pred=1	total	Recall
Ans=0	18,930	6,111	25,041	0.7702
Ans=1	3,521	11,438	14,959	0.777525
total	22,451	17,549	40,000	
Precision	0.8495	0.663325		
F1 score	0.7948	0.68705		
Accuracy	0.7592			

Large network Rebinarized

	Pred=0	Pred=1	Total	Recall
Ans=0	29,179	1,486	30,665	0.949625
Ans=1	1,984	7,351	9,335	0.751825
Total	31,163	8,837	40,000	
Precision	0.924925	0.739125		
F1 score	0.9362	0.7347		
Accuracy	0.91325			

Figure 4. K-Fold Cross-Validation

(A) K-fold cross-validation based on four independent datasets, of which three were used for training and one was used for validation, in all possible combinations.

(B) K-fold cross-validation was performed using the small network trained on automatically binarized target blocks, and using the large network trained on rebinarized target blocks. Data are macro averaged F1 score and accuracy.

(C) Detailed K-fold cross-validation data.

See also Table S4.

Matrigel, and routinely passaged every week. Media were changed every other day. Detailed protocols are described in Supplemental Experimental Procedures.

Endothelial Cell Differentiation

iPSCs cultured on Matrigel-coated 6-well plates were enzymatically detached on day 7, and differentiated into endothelial cells as described in Supplemental Experimental Procedures.

Flow Cytometry

At day 6 of differentiation, cells were dissociated, stained with APC-conjugated anti-CD31, and sorted on BD FACSAria III. As a negative control, we used unstained cells. Detailed protocols are described in Supplemental Experimental Procedures.

Immunocytochemistry

At day 6 of differentiation, cells were fixed with 4% paraformaldehyde, blocked with ImmunoBlock, probed with primary antibodies to CD31, and labeled with secondary antibodies as described in Supplemental Experimental Procedures.

Preparation of Datasets

All phase-contrast and immunofluorescent images were acquired at day 6 of differentiation. Two hundred images were automatically obtained from each of four independent experiments. Of these, 640 were used for training and 160 were used for validation in Figures 2 and 3. For K-fold validation in Figure 4, 600 images from three experiments were used for training and 200 images from one experiment were used for validation, in all possible combinations. Datasets were constructed by randomly extracting 200 input blocks from each phase-contrast image. On the other hand, target blocks were extracted from binarized immunofluorescent images. Detailed procedures are described in Supplemental Experimental Procedures.

Deep Neural Networks

We used LeNet, a small network that contains two convolution layers, two max pooling layers, and two fully connected layers, as well as AlexNet, a large network that contains five convolution layers, three max pooling layers, and three fully connected layers. Network structures are shown in Figure 3A and Supplemental Experimental Procedures.

Performance Evaluation

Performance was evaluated based on F1 scores, an aggregate of recall and precision, and on accuracy, the fraction of correct predictions. Detailed information is provided in Supplemental Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at https://doi.org/10.1016/j.stemcr.2018.04.007.

AUTHOR CONTRIBUTIONS

D.K., T. Kunihiro, and S.Y. designed experiments. D.K., M.L., T. Kunihiro, S.Y., Y.K., M.K., T. Katsuki, S.I., T.S., and K.F. collected data.

D.K., M.L., and T. Kunihiro analyzed data. K.F. supervised the research. D.K. and S.Y. wrote the article.

ACKNOWLEDGMENTS

We thank all of our laboratory members for assistance. This research was supported by Grants-in-Aid for Scientific Research (JSPS KAKENHI grant numbers 16H05304 and 16K15415), by SENSHIN Medical Research Foundation, by Suzuken Memorial Foundation, and by Keio University Medical Science Fund. K.F. is a co-founder of and has equity in Heartseed. T.K. is an employee of Sony Imaging Products & Solutions.

Received: December 4, 2017
Revised: April 13, 2018
Accepted: April 13, 2018
Published: May 10, 2018

REFERENCES

Adams, R.H., and Alitalo, K. (2007). Molecular regulation of angiogenesis and lymphangiogenesis. Nat. Rev. Mol. Cell Biol. *8*, 464–478.

Avior, Y., Sagi, I., and Benvenisty, N. (2016). Pluripotent stem cells in disease modelling and drug discovery. Nat. Rev. Mol. Cell Biol. *17*, 170–182.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2006). Greedy layer-wise training of deep networks. In Proceedings of the 19th International Conference on Neural Information Processing Systems, B. Schölkopf, J.C. Platt, and T. Hoffman, eds. (MIT Press), pp. 153–160.

Buggenthin, F., Buettner, F., Hoppe, P.S., Endele, M., Kroiss, M., Strasser, M., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K.D., Hilsenbeck, O., et al. (2017). Prospective identification of hematopoietic lineage choice by deep learning. Nat. Methods *14*, 403–406.

Chen, I.Y., Matsa, E., and Wu, J.C. (2016). Induced pluripotent stem cells: at the heart of cardiovascular precision medicine. Nat. Rev. Cardiol. *13*, 333–349.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.

Gorodeski, E.Z., Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Hsich, E., Zhang, Z.M., Vitolins, M.Z., Manson, J.E., Curb, J.D., Martin, L.W., et al. (2011). Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the Women's Health Initiative. Circ. Cardiovasc. Qual. Outcomes *4*, 521–532.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA *316*, 2402–2410.

Hatipoglu, N., and Bilgin, G. (2014). Classification of histopathological images using convolutional neural network. Paper presented at: 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA).

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. https://doi.org/10.1109/CVPR.2016.90.

Heylman, C., Datta, R., Sobrino, A., George, S., and Gratton, E. (2015). Supervised machine learning for classification of the electrophysiological effects of chronotropic drugs on human induced pluripotent stem cell-derived cardiomyocytes. PLoS One *10*, e0144572.

Hinton, G.E., Osindero, S., and Teh, Y.W. (2006). A fast learning algorithm for deep belief nets. Neural Comput. *18*, 1527–1554.

Hsich, E., Gorodeski, E.Z., Blackstone, E.H., Ishwaran, H., and Lauer, M.S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circ. Cardiovasc. Qual. Outcomes *4*, 39–45.

Hu, B.Y., Weick, J.P., Yu, J., Ma, L.X., Zhang, X.Q., Thomson, J.A., and Zhang, S.C. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. Proc. Natl. Acad. Sci. USA *107*, 4335–4340.

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. ArXiv. https://arxiv.org/pdf/1502.03167.pdf.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, eds. (Curran Associates Inc.), pp. 1097–1105.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE *86*, 2278–2324.

Lindblom, P., Gerhardt, H., Liebner, S., Abramsson, A., Enge, M., Hellstrom, M., Backstrom, G., Fredriksson, S., Landegren, U., Nystrom, H.C., et al. (2003). Endothelial PDGF-B retention is required for proper investment of pericytes in the microvessel wall. Genes Dev. *17*, 1835–1840.

McCulloch, W.S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biol. *5*, 115–133.

Niioka, H., Asatani, S., Yoshimura, A., Ohigashi, H., Tagawa, S., and Miyake, J. (2018). Classification of C2C12 cells at differentiation by convolutional neural network of deep learning using phase contrast images. Hum. Cell *31*, 87–93.

Osafune, K., Caron, L., Borowiak, M., Martinez, R.J., Fitz-Gerald, C.S., Sato, Y., Cowan, C.A., Chien, K.R., and Melton, D.A. (2008). Marked differences in differentiation propensity among human embryonic stem cell lines. Nat. Biotechnol. *26*, 313–315.

Patsch, C., Challet-Meylan, L., Thoma, E.C., Urich, E., Heckel, T., O'Sullivan, J.F., Grainger, S.J., Kapp, F.G., Sun, L., Christensen, K., et al. (2015). Generation of vascular endothelial and smooth muscle cells from human pluripotent stem cells. Nat. Cell Biol. *17*, 994–1003.

Saha, S., and Vemuri, R. (2000). An analysis on the effect of image activity on lossy coding performance. Paper presented at: 2000 IEEE International Symposium on Circuits and Systems Emerging Technologies for the 21st Century Proceedings (IEEE Cat No 00CH36353).

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. ArXiv. https://arxiv.org/pdf/1409.1556.pdf.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. ArXiv. https://arxiv.org/pdf/1409.4842.pdf.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Takakura, N., Watanabe, T., Suenobu, S., Yamada, Y., Noda, T., Ito, Y., Satake, M., and Suda, T. (2000). A role for hematopoietic stem cells in promoting angiogenesis. Cell *102*, 199–209.

Tanaka, A., Yuasa, S., Mearini, G., Egashira, T., Seki, T., Kodaira, M., Kusumoto, D., Kuroda, Y., Okata, S., Suzuki, T., et al. (2014). Endothelin-1 induces myofibrillar disarray and contractile vector variability in hypertrophic cardiomyopathy-induced pluripotent stem cell-derived cardiomyocytes. J. Am. Heart Assoc. *3*, e001263.

Van Valen, D.A., Kudo, T., Lane, K.M., Macklin, D.N., Quach, N.T., DeFelice, M.M., Maayan, I., Tanouchi, Y., Ashley, E.A., and Covert, M.W. (2016). Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. PLoS Comput. Biol. *12*, e1005177.

Yuan-Hsiang, C., Abe, K., Yokota, H., Sudo, K., Nakamura, Y., Cheng-Yu, L., and Ming-Dar, T. (2017). Human induced pluripotent stem cell region recognition in microscopy images using convolutional neural networks. Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2017, 4058–4061.

Yuasa, S., and Fukuda, K. (2008). Cardiac regenerative medicine. Circ. J. *72* (*Suppl A*), A49–A55.

Zeng, X., Ouyang, W., Yan, J., Li, H., Xiao, T., Wang, K., Liu, Y., Zhou, Y., Yang, B., Wang, Z., et al. (2016). Crafting GBD-Net for object detection. ArXiv. https://arxiv.org/pdf/1610.02579.pdf.

Stem Cell Reports, Volume *10*

Supplemental Information

Automated Deep Learning-Based System to Identify Endothelial Cells Derived from Induced Pluripotent Stem Cells

Dai Kusumoto, Mark Lachmann, Takeshi Kunihiro, Shinsuke Yuasa, Yoshikazu Kishino, Mai Kimura, Toshiomi Katsuki, Shogo Itoh, Tomohisa Seki, and Keiichi Fukuda

Supplemental Figure Legends

Figure S1. Generation of iPSC-derived Endothelial Cells

(A) Differentiation of endothelial cells. iPSCs were seeded onto Matrigel-coated dishes, cultured in indicated conditions, and examined at day 6.

(B) Phase-contrast images at day 1 to 6. Scale bars, 500 μm.

(C) Phase-contrast images (upper panels), immunofluorescent staining for CD31 (middle panels), and FACS analysis (bottom panels) showed variability in differentiation at day 6 in various experiments. Left, middle, and right panels show experiments with high, intermediate, and low differentiation efficiency. Scale bars, 200 μm.

Figure S2. Network Performance Depending on Input Block Size, Staining Threshold and Target Block Size, Related to Figure 2, Tables S1 and S2.

(A) Phase-contrast and binarized fluorescent images of 512 × 512 px, 256 × 256 px, 128 × 128 px, 64 × 64 px, and 32 × 32 px blocks. Scale bars, 80 μm, 40 μm, 20 μm, 10 μm, and 5 μm, respectively.

(B) and (C) Accuracy obtained from networks trained on 32 × 32 px, 64 × 64 px, 128 × 128 px, 256 × 256 px, and 512 × 512 px input blocks, using 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 as staining threshold, *i.e.*, the ratio of white pixels to black for a binarized image to be classified as stained..

(D) F1 score and accuracy obtained from networks trained on input and target blocks of various sizes.

Figure S3. Optimization of Network Performance, Related to Figure 3.

(A) Learning curve of the small and large network, as assessed by accuracy.

(B) Representative images of true positives and true negatives (blue) and of false positives and false negatives (red). Yellow areas are CD31-stained. Scale bars, 200 μm.

Figure S4. Correlation Between Image Complexity and F1 score, Related to Figure 3.

(A) Representative phase-contrast images with complexity 0.00-0.04 (group 1), 0.04-0.08 (group 2), and over 0.08 (group 3). Scale bars, 80 μm.

(B) F1 score in each group using the small and large network (left), and relationship between F1 score and image complexity (right).

(C) and (D) Performance statistics from each group (C) and over increasing complexity (D). True positive: TP, True negative: TN, False positive: FP, and False negative: FN

(E) Time required to classify each block.

Supplemental Table Legends

Table S1. Number of Blocks Required for Learning, Related to Figure 2.
Networks were trained on 500, 1,000, 2,000, 4,000, 8,000, 16,000, 32,000, 64,000, and 128,000 blocks. Accuracy, recall, precision, and F1 score were assessed using 128 × 128 px input blocks and 128 × 128 px target blocks (left), or using 512 × 512 px input blocks and 32 × 32 px target blocks (right).

Table S2. Network Performance Depending on Input Block Size and Staining Threshold, Related to Figures 2 and S2.
Networks were trained using 32 × 32 px, 64 × 64 px, 128 × 128 px, 256 × 256 px, and 512 × 512 px input blocks, using 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 as staining threshold, *i.e.*, the ratio of white pixels to black for a binarized image to be classified as stained. Accuracy, recall, precision, and F1 score were calculated.

Table S3. Network Performance Depending on Target Block Size, Related to Figures 2 and S2.
F1 score, accuracy and other indices obtained from networks trained on input and target blocks of various sizes.

Table S4. Network Performance, Related to Figures 3 and 4.
(A) Network performance was compared following training on automatically binarized or rebinarized fluorescent images.
(B) K-fold cross validation of the small network trained on automatically binarized fluorescent images (left), and of the large network trained on rebinarized fluorescent images (right). Independent training and validation were performed according to Figure 4.

Supplemental Experimental Procedures

iPSC Culture

iPSCs were maintained in mTeSR1 (Stem Cell Technologies, Vancouver, BC, Canada) media with 0.5 % penicillin/streptomycin (Thermo Fisher Scientific, Waltham, MA, USA) on culture dishes coated with growth factor-reduced Matrigel (BD Biosciences, San Jose, CA, USA). iPSCs were routinely passaged every week by washing in PBS, incubating in TrypLE Select (Thermo Fisher Scientific) for 3 min at 37 °C, detaching with a cell scraper, harvesting, and reseeding at a split ratio of 1:5 to 1:8 in mTeSR1 with 0.5 % penicillin/streptomycin and 10 μM ROCK inhibitor Y-27632 (Wako, Osaka, Japan). Media were changed every other day.

Endothelial Cell Differentiation

iPSCs cultured on Matrigel-coated 6-well plates were detached using TrypLE Select on day 7, and clumps with diameter 100-200 μm were reseeded on Matrigel-coated dishes and incubated for 24 hours in mTeSR1 media with 10 μM ROCK inhibitor Y-27632. On day 1, mesoderm was induced in N2B27 media (1:1 mixture of DMEM/F12 and Neurobasal media containing N2 and B27, all reagents from Thermo Fisher Scientific) supplemented with β-mercaptoethanol, 8 μM CHIR-99021 (Cayman Chemical, Ann Arbor, MI, USA), and 25 ng/mL BMP4 (R&D Systems, Minneapolis, MN, USA). At day 3 and 4, media were replaced with StemPro-34 SFM (Thermo Fisher Scientific) containing 200 ng/mL VEGF (Wako) and 2 μM forskolin (Abcam, Cambridge, UK) to induce endothelial cell specification (Patsch et al., 2015). Endothelial cell clusters were reliably obtained on day 6. After sorting by flow cytometry, cells expressing CD31 were cultured for another four days in StemPro-34 SFM containing 50 ng/mL VEGF.

Flow Cytometry

At day 6 of differentiation, cells were dissociated into singe cells using Accutase (Innovative Cell Technologies, San Diego, CA, USA), suspended in PBS with 0.5 % BSA, and stained with a 1:50 dilution of APC-conjugated anti-CD31 (Miltenyi Biotec, Bergisch Gladbach, NRW, Germany, catalog no. 130-092-652) according to the manufacturer's instructions. As a negative control, we used unstained cells. Cells were then sorted on a BD FACS Aria III (Becton Dickinson, Franklin Lakes, NJ, USA), and data were collected from at least 10,000 events.

Immunocytochemistry

Cells were fixed in 4 % paraformaldehyde (MUTO Pure Chemicals, Tokyo, Japan) for 20 min at room temperature, washed with PBS, blocked with ImmunoBlock (DS Pharma Biomedical, Osaka, Japan) for 1 h, and probed at 4 °C overnight with 1:20 primary antibodies to CD31 (R&D Systems, catalog no. AF806). Specimens were then washed thrice in PBS, labeled for 1 h with 1:200 secondary anti-sheep IgG (Thermo Fisher Scientific, catalog no. A-11015), and imaged on an inverted fluorescence phase-contrast microscope.

Preparation of Datasets

Phase-contrast and immunofluorescent images were acquired at day 6 of differentiation. Two hundred images were automatically acquired from each of four independent experiments. Phase contrast and fluorescent images were taken on an SI8000 Research Microscope (SONY, Tokyo, Japan) at 10× and 0.454 μm/pixel. Each image was saved as a

2752 × 2200 px grayscale image in BMP format at 8 bits per pixel. To generate datasets for training and evaluation, 200 input blocks of 32 × 32 px, 64 × 64 px, 128 × 128 px, 256 × 256 px, and 512 × 512 px were randomly extracted from each phase-contrast image. The 256 × 256 px and 512 × 512 px input blocks were resized to 128 × 128 px as needed. Immunofluorescent images of CD31 were binarized using in-house software to distinguish specific signals from nonspecific signals. In particular, pixels were binarized to white if its value (0-255 in raw immunofluorescent images) is above a threshold value empirically determined based on the complete image. All other pixels were binarized to black. Finally, 32 × 32 px and 128 × 128 px target blocks were extracted, corresponding to the center of input blocks.

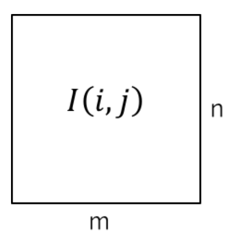
Data in Figure 2 and 3 were generated based on 640 training images and 160 validation images. In both experiments in Figure 2A, 500, 1,000, 2,000, 4,000, 8,000, 16,000, 32,000, 64,000, and 128,000 blocks were used for training, and 32,000 blocks were used for validation. In Figure 2B, 32,000 blocks were used for training, and 32,000 blocks were used for validation. In Figure 2C to 3E, all 128,000 blocks were used for training, and 32,000 blocks were used for validation. For K-fold validation in Figure 4, four independent data sets of 200 images each were obtained, of which three were used as training sets and one was used as validation set in all possible combinations, such that the number of folds is 4. To rebinarize target blocks, we compared raw fluorescent images to phase-contrast images in GNU Image Manipulation Program, and rebinarized weakly stained, dense colonies as black pixels. All 800 images were processed in this manner.

Deep Neural Networks

We used LeNet, a small convolutional neural network with two convolution layers, two max pooling layers, and two fully-connected layers, as well as AlexNet, a large network with five convolution layers, three max pooling layers, and three fully-connected layers (Figure 3A). In both networks, each convolutional layer is connected to Rectified Linear Units for activation (Nair and Hinton, 2010). In the output layer, we used a sigmoid function, consistent with binary classification. We used mini-batch training with stochastic gradient descent, learning rate 0.01, cross-entropy error as loss function. Weights were initialized using the Xavier algorithm (Glorot and Bengio, 2010). To avoid overfitting, dropout techniques were used in the large network. Networks were trained using the TensorFlow/Keras framework (Cholle, 2015) on a computer with a Core i7-6700 CPU (Intel, Santa Clara, CA, USA), 16 GB memory, and GeForce GTX980Ti GPU (NVIDIA, Santa Clara, CA, USA).

Image Complexity

We calculated image complexity (activity), which we used as an index of cell density, in all 32,000 512 × 512 px validation blocks used in the small and large network. This value was



$$Activity = \frac{\sum_{i=0}^{m-2}\sum_{j=0}^{n-1}|I(i,j)-I(i+1,j)| + \sum_{i=0}^{m-1}\sum_{j=0}^{n-2}|I(i,j)-I(i,j+1)|}{\{(\mathrm{m}-1)\mathrm{n}+\mathrm{m}(\mathrm{n}-1)\}\{\max(I)-\min(I)\}}$$

calculated according to

where m is the image width in pixels, n is the image height in pixels, *I* is the pixel value, and *(i, j)* are coordinates (x-axis, y-axis). Essentially, image complexity is the average difference between adjacent pixels normalized to the dynamic range (Saha and Vemuri, 2000). Thus, the numerator is the sum of differences in adjacent pixels on both x and y axes, while the denominator is the product of image size and dynamic range, which is the difference between the maximum and minimum pixel value.

Evaluation of Prediction Performance

Network performance was evaluated based on accuracy and F1 score, which combines recall (sensitivity) and precision (true positive rate). Accordingly, the F1 score is 1 for perfect predictions and 0.5 for random predictions. On the other hand, precision is the fraction of true positives among predicted positives, while recall is the fraction of true positives detected among all positives:

$$F1\ score\ = \frac{2Recall\ \times\ Precision}{Recall + Precision}, Precision\ = \frac{TP}{TP + FP}, Recall\ = \frac{TP}{TP + FN}$$

Precision and recall for negative predictions were calculated in a similar manner:

$$Precision\ (negative) = \frac{TN}{TN + FN}, Recall\ (negative)\ = \frac{TN}{TN + FP}$$

Finally, accuracy is the ratio of correct predictions to all predictions:

$$Accuracy = \frac{TP + TN}{TP + FP + TN +\ FP}$$

Supplemental References

Cholle, F. (2015). Keras (https://github.com/fchollet/keras: GitHub).

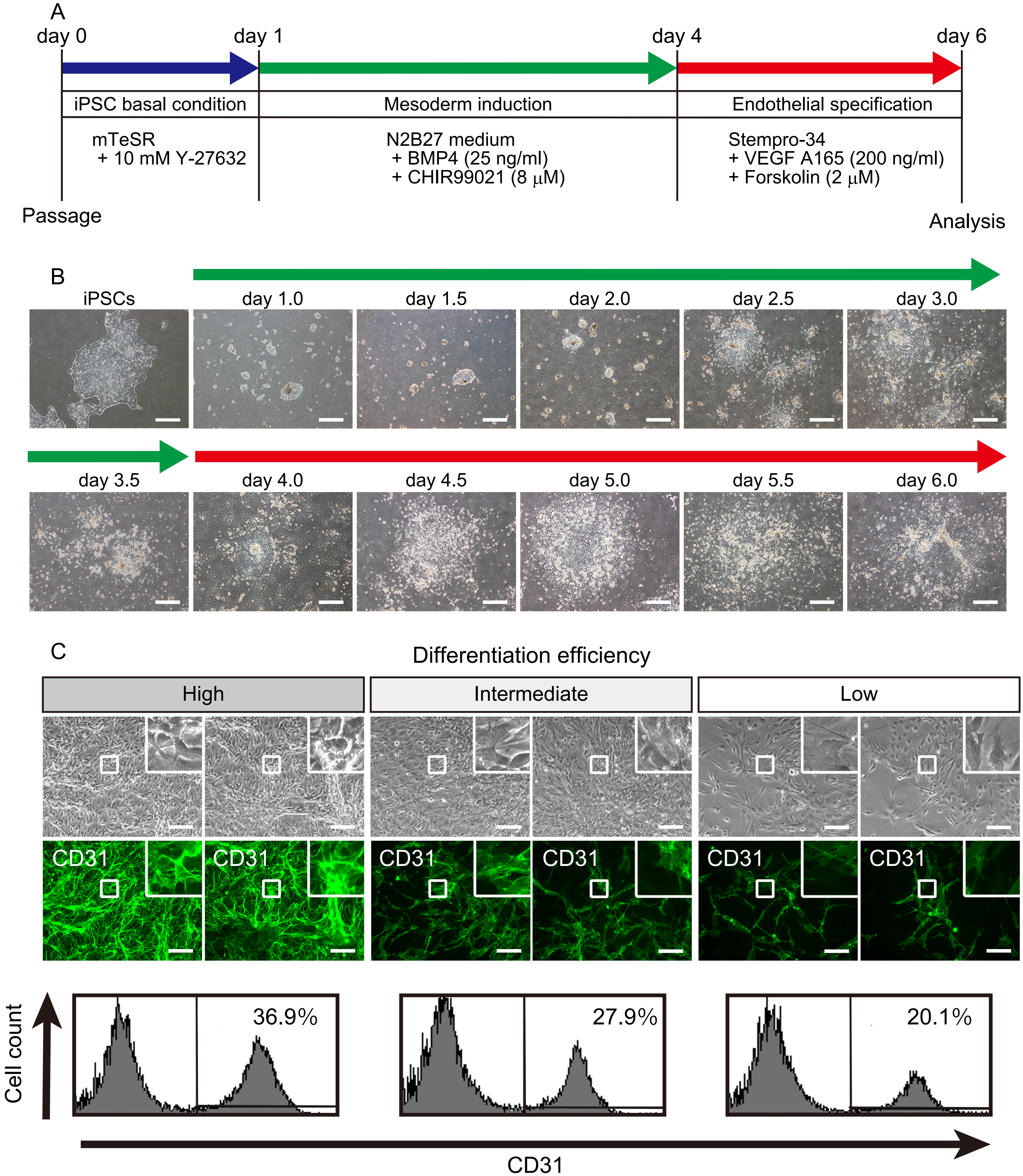
Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, T. Yee Whye, and T. Mike, eds. (Proceedings of Machine Learning Research: PMLR), pp. 249--256.

Nair, V., and Hinton, G.E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning (Haifa, Israel: Omnipress), pp. 807-814.

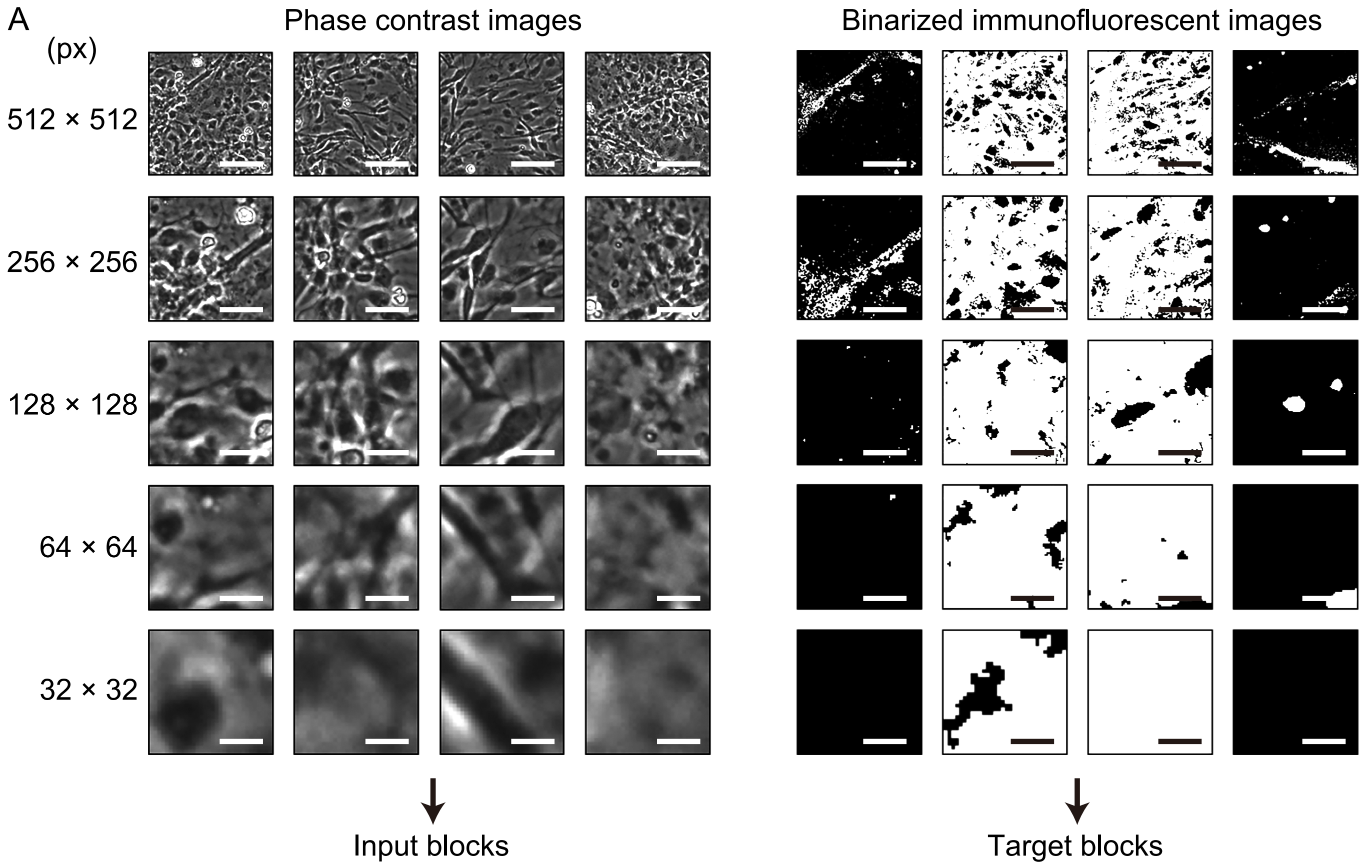
Patsch, C., Challet-Meylan, L., Thoma, E.C., Urich, E., Heckel, T., O'Sullivan, J.F., Grainger, S.J., Kapp, F.G., Sun, L., Christensen, K., *et al.* (2015). Generation of vascular endothelial and smooth muscle cells from human pluripotent stem cells. Nat Cell Biol *17*, 994-1003.

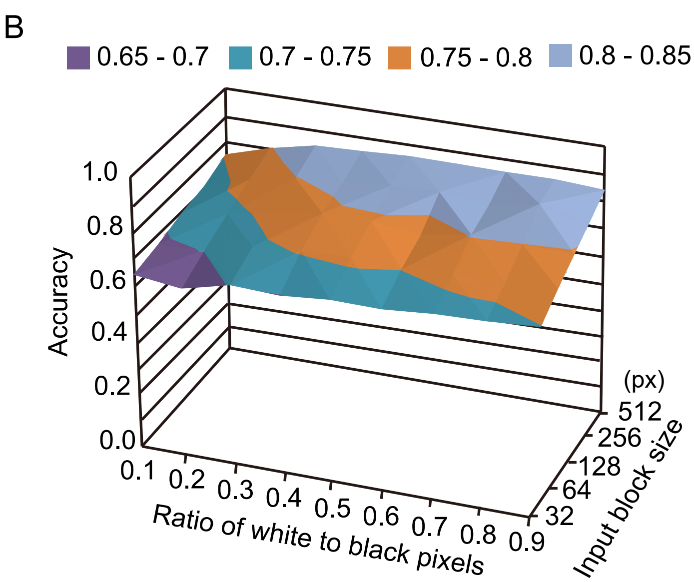
Saha, S., and Vemuri, R. (2000). An analysis on the effect of image activity on lossy coding performance. Paper presented at: 2000 IEEE International Symposium on Circuits and Systems Emerging Technologies for the 21st Century Proceedings (IEEE Cat No00CH36353).

Supplementary Figure 1



Supplementary Figure 2



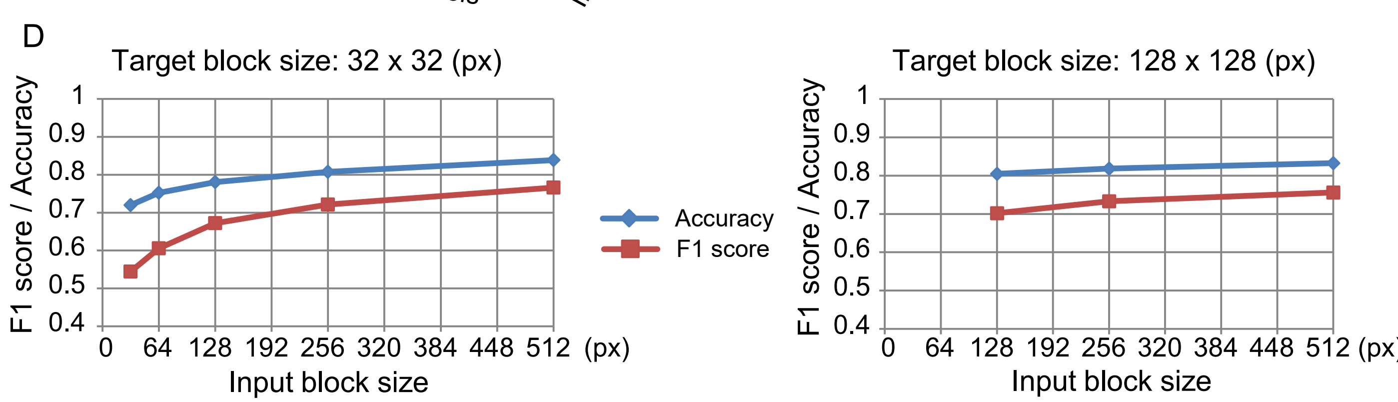


C

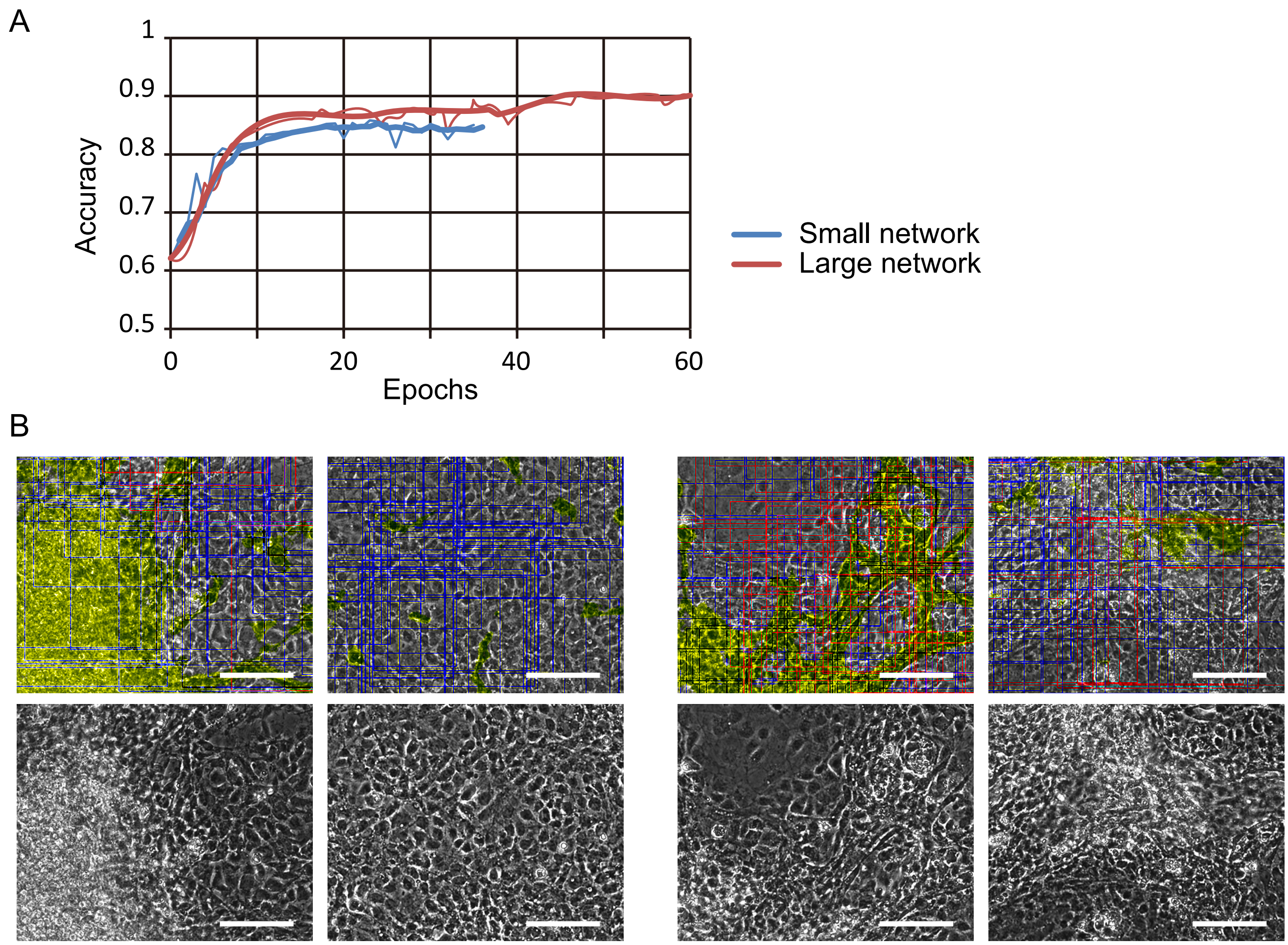
Accuracy

Input block size (px)

Ratio of white to black pixels	32	64	128	256	512	Average
0.1	0.6786	0.6919	0.7121	0.7230	0.7529	0.7117
0.2	0.6624	0.7019	0.7364	0.7685	0.7978	0.7334
0.3	0.7028	0.7389	0.7626	0.7988	0.8304	0.7667
0.4	0.6986	0.7546	0.7900	0.8084	0.8366	0.7776
0.5	0.7144	0.7599	0.7925	0.8258	0.8448	0.7875
0.6	0.7143	0.7509	0.7745	0.8304	0.8444	0.7829
0.7	0.7295	0.7635	0.8027	0.8140	0.8431	0.7906
0.8	0.7364	0.7639	0.7994	0.8448	0.8448	0.7979
0.9	0.7491	0.7752	0.7954	0.8147	0.8360	0.7941
Average	0.7096	0.7445	0.7740	0.8032	0.8256	

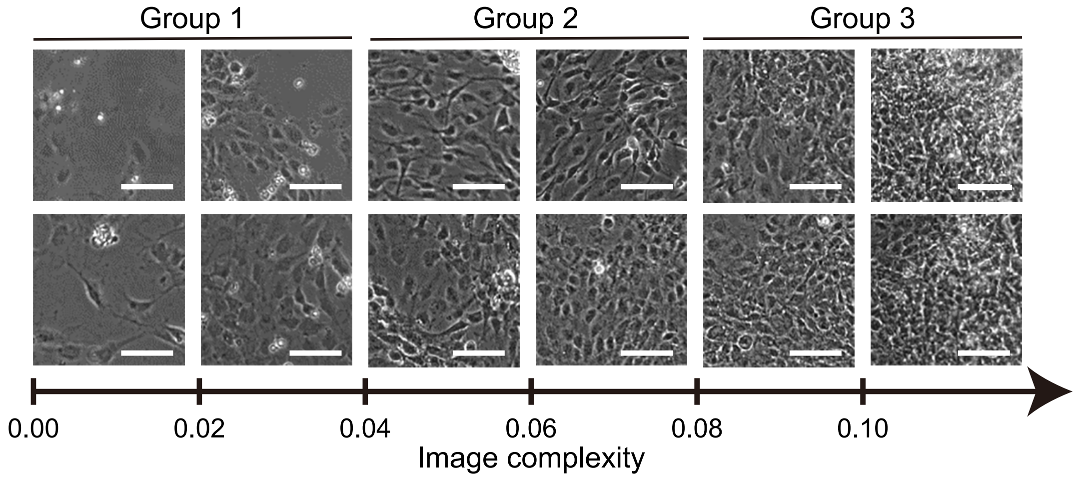


Supplementary Figure 3

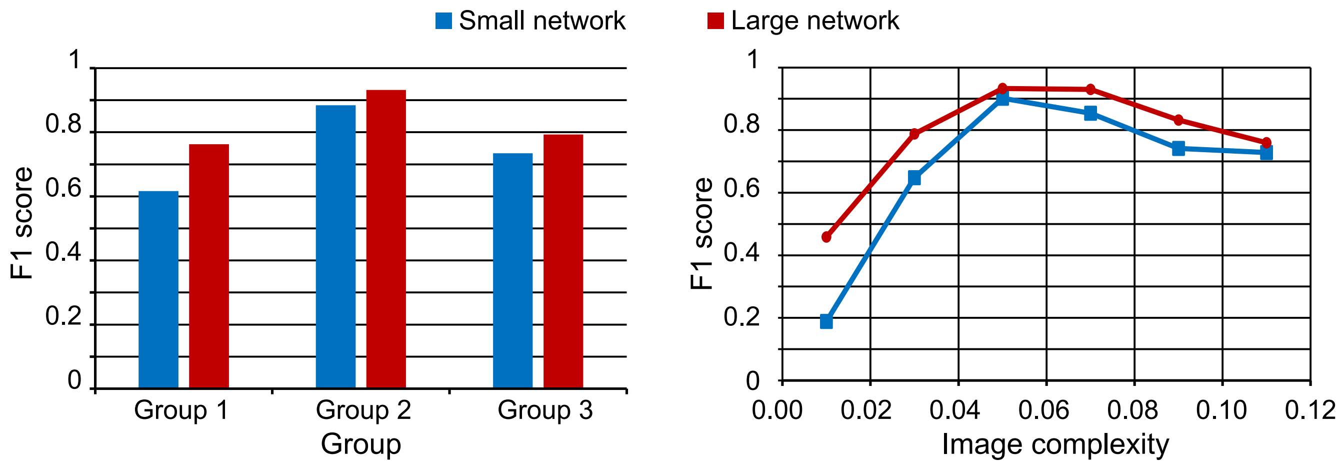


Supplementary Figure 4

A



B



C

Small network

Group	Image complexity	TN	FN	FP	TP	Precision	Recall	Accuracy	F1 score
1	0.00-0.04	8,657	707	546	1,006	0.648196	0.587274	0.885214	0.616233
2	0.04-0.08	2,758	992	633	6,192	0.907253	0.861915	0.846336	0.884003
3	0.08-	6,628	1,025	604	2,252	0.788515	0.687214	0.84499	0.734388

Large network

Group	Image complexity	TN	FN	FP	TP	Precision	Recall	Accuracy	F1 score
1	0.00-0.04	8,870	452	333	1,261	0.791092	0.736135	0.928087	0.762625
2	0.04-0.08	2,818	413	573	6,771	0.921977	0.942511	0.906761	0.932131
3	0.08-	6,603	709	629	2,568	0.803253	0.783644	0.872681	0.793327

D

Small network

Image complexity	TN	FN	FP	TP	Precision	Recall	Accuracy	F1 score
0.00-0.02	4,595	165	16	21	0.567568	0.112903	0.962268	0.188341
0.02-0.04	4,062	542	530	985	0.650165	0.645056	0.824808	0.6476
0.04-0.06	1,105	475	395	3,991	0.909941	0.893641	0.854174	0.901717
0.06-0.08	1,653	517	238	2,201	0.902419	0.809787	0.83619	0.853597
0.08-0.10	3,078	487	219	1,014	0.822384	0.67555	0.852855	0.74177
0.10-	3,550	538	385	1,238	0.762785	0.697072	0.838382	0.72845

Large network

Image complexity	TN	FN	FP	TP	Precision	Recall	Accuracy	F1 score
0.00-0.02	4,602	128	9	58	0.865672	0.311828	0.97144	0.458498
0.02-0.04	4,268	324	324	1,203	0.787819	0.787819	0.8941	0.787819
0.04-0.06	1,149	251	351	4,215	0.923127	0.943798	0.899095	0.933348
0.06-0.08	1,669	162	222	2,556	0.920086	0.940397	0.916685	0.930131
0.08-0.10	3,039	247	258	1,254	0.829365	0.835443	0.894748	0.832393
0.10-	3,564	462	371	1,314	0.779822	0.739865	0.854141	0.759318

E

Small network: 100 μsec
Large network: 275 μsec

Supplementary Table 1

Pred = “Prediction” ; Ans = “Answer” ; 0 = “unstained” ; 1 = “stained”

Number of input blocks

Input block size: 128 x 128 (px)
Target block size: 128 x 128 (px)

500

	Pred=0	Pred=1	Total	Recall
Ans=0	21,261	0	21,261	1
Ans=1	10,739	0	10,739	0
Total	32,000	0	32,000	
Precision	0.6644	0		
F1 score	0.7984	0		
Accuracy	0.6644			

1,000

	Pred=0	Pred=1	Total	Recall
Ans=0	21,261	0	21,261	1
Ans=1	10,737	2	10,739	0.0002
Total	31,998	2	32,000	
Precision	0.6644	1		
F1 score	0.7984	0.0004		
Accuracy	0.6645			

2,000

	Pred=0	Pred=1	Total	Recall
Ans=0	21,261	0	21,261	1
Ans=1	10,738	1	10,739	0.0001
Total	31,999	1	32,000	
Precision	0.6644	1		
F1 score	0.7984	0.0002		
Accuracy	0.6644			

4,000

	Pred=0	Pred=1	Total	Recall
Ans=0	20,913	348	21,261	0.9836
Ans=1	9,901	838	10,739	0.078
Total	30,814	1,186	32,000	
Precision	0.6787	0.7066		
F1 score	0.8032	0.1405		
Accuracy	0.6797			

8,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,933	2,328	21,261	0.8905
Ans=1	6,646	4,093	10,739	0.3811
Total	25,579	6,421	32,000	
Precision	0.7402	0.6374		
F1 score	0.8084	0.477		
Accuracy	0.7196			

16,000

	Pred=0	Pred=1	Total	Recall
Ans=0	19,298	1,963	21,261	0.9077
Ans=1	5,765	4,974	10,739	0.4632
Total	25,063	6,937	32,000	
Precision	0.77	0.717		
F1 score	0.8332	0.5628		
Accuracy	0.7585			

32,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,649	2,612	21,261	0.8771
Ans=1	4,635	6,104	10,739	0.5684
Total	23,284	8,716	32,000	
Precision	0.8009	0.7003		
F1 score	0.8373	0.6275		
Accuracy	0.7735			

64,000

	Pred=0	Pred=1	Total	Recall
Ans=0	19,007	2,254	21,261	0.894
Ans=1	4,019	6,720	10,739	0.6258
Total	23,026	8,974	32,,000	
Precision	0.8255	0.7488		
F1 score	0.8584	0.6818		
Accuracy	0.804			

128,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,409	2,852	21,261	0.8659
Ans=1	3,388	7,351	10,739	0.6845
Total	21,797	10,203	32,000	
Precision	0.8446	0.7205		
F1 score	0.8551	0.702		
Accuracy	0.805			

Input block size: 512 x 512 (px)
Target block size: 32 x 32 (px)

500

	Pred=0	Pred=1	Total	Recall
Ans=0	21,148	0	21,148	1
Ans=1	10,852	0	10,852	0
Total	32,000	0	32,000	
Precision	0.6609	0		
F1 score	0.7958	0		
Accuracy	0.6609			

1,000

	Pred=0	Pred=1	Total	Recall
Ans=0	21,148	0	21,148	1
Ans=1	10,852	0	10,852	0
Total	32,000	0	32,000	
Precision	0.6609	0		
F1 score	0.7958	0		
Accuracy	0.6609			

2,000

	Pred=0	Pred=1	Total	Recall
Ans=0	21,129	19	21,148	0.9991
Ans=1	10,774	78	10,852	0.0072
Total	31,903	97	32,000	
Precision	0.6623	0.8041		
F1 score	0.7966	0.0142		
Accuracy	0.6627			

4,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,684	2,464	21,148	0.8835
Ans=1	8,276	2,576	10,852	0.2374
Total	26,960	5,040	32,000	
Precision	0.693	0.5111		
F1 score	0.7768	0.3242		
Accuracy	0.6644			

8,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,836	2,312	21,148	0.8907
Ans=1	6,514	4,338	10,852	0.3997
Total	25,350	6,650	32,000	
Precision	0.743	0.6523		
F1 score	0.8102	0.4957		
Accuracy	0.7242			

16,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,816	2,332	21,148	0.8897
Ans=1	4,113	6,739	10,852	0.621
Total	22,929	9,071	32,000	
Precision	0.8206	0.7429		
F1 score	0.8538	0.6765		
Accuracy	0.7986			

32,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,743	2,405	21,148	0.8863
Ans=1	3,865	6,987	10,852	0.6438
Total	22,608	9,392	32,000	
Precision	0.829	0.7439		
F1 score	0.8567	0.6903		
Accuracy	0.8041			

64,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,754	2,394	21,148	0.8868
Ans=1	3,225	7,627	10,852	0.7028
Total	21,979	10,021	32,000	
Precision	0.8533	0.7611		
F1 score	0.8697	0.7308		
Accuracy	0.8244			

128,000

	Pred=0	Pred=1	Total	Recall
Ans=0	18,558	2,590	21,148	0.8775
Ans=1	2,730	8,122	10,852	0.7484
Total	21,288	10,712	32,000	
Precision	0.8718	0.7582		
F1 score	0.8746	0.7533		
Accuracy	0.8337			

Supplementary Table 2

Pred = “Prediction” : 0 = “unstained” , 1 = “stained”
Ans = “Answer” : 0 = “unstained” , 1 = “stained”

Input block size (px)

Ratio of white to black pixels

0.1

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	16,467	2,885	19,352	0.8509	Ans=0	14,370	4,256	18,626	0.7715	Ans=0	13,616	4,235	17,851	0.7628	Ans=0	14,587	2,392	16,979	0.8591	Ans=0	13,447	2,771	16,218	0.8291
Ans=1	7,401	5,247	12,648	0.4148	Ans=1	5,602	7,772	13,374	0.5811	Ans=1	4,979	9,170	14,149	0.6481	Ans=1	6,471	8,550	15,021	0.5692	Ans=1	5,137	10,645	15,782	0.6745
Total	23,868	8,132	32,000		Total	19,972	120,28	32,000		Total	18,595	13,405	32,000		Total	21,058	10,942	32,000		Total	18,584	13,416	32,000	
Precision	0.6899	0.6452			Precision	0.7195	0.6462			Precision	0.7322	0.6841			Precision	0.6927	0.7814			Precision	0.7236	0.7935		
F1 score	0.762	0.505			F1 score	0.7446	0.6119			F1 score	0.7472	0.6656			F1 score	0.767	0.6586			F1 score	0.7728	0.7292		
Accuracy	0.6786				Accuracy	0.6919				Accuracy	0.7121				Accuracy	0.723				Accuracy	0.7529			

0.2

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	13,946	5,999	19,945	0.6992	Ans=0	14,719	4,817	19,536	0.7534	Ans=0	15,570	3,594	19,164	0.8125	Ans=0	16,199	2,723	18,922	0.8561	Ans=0	15,390	3,067	18,457	0.8338
Ans=1	4,805	7,250	12,055	0.6014	Ans=1	4,721	7,743	12,464	0.6212	Ans=1	4,840	7,996	12,836	0.6229	Ans=1	4,685	8,393	13,078	0.6418	Ans=1	3,403	10,140	13,543	0.7487
Total	18,751	13,249	32,000		Total	19,440	12,560	32,000		Total	20,410	11,590	32,000		Total	20,884	11,116	32,000		Total	18,793	13,207	32,000	
Precision	0.7437	0.5472			Precision	0.7572	0.6165			Precision	0.7629	0.6899			Precision	0.7757	0.755			Precision	0.8189	0.7678		
F1 score	0.7208	0.573			F1 score	0.7553	0.6188			F1 score	0.7869	0.6547			F1 score	0.8139	0.6938			F1 score	0.8263	0.7581		
Accuracy	0.6624				Accuracy	0.7019				Accuracy	0.7364				Accuracy	0.7685				Accuracy	0.7978			

0.3

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	16,304	4,081	20,385	0.7998	Ans=0	17,128	3,042	20,170	0.8492	Ans=0	17,085	2,975	20,060	0.8517	Ans=0	17,001	3,035	20,036	0.8485	Ans=0	17,963	1,863	19,826	0.906
Ans=1	5,430	6,185	11,615	0.5325	Ans=1	5,314	6,516	11,830	0.5508	Ans=1	4,623	7,317	11,940	0.6128	Ans=1	3,404	8,560	11,964	0.7155	Ans=1	3,564	8,610	12,174	0.7072
Total	21,734	10,266	32,000		Total	22,442	9,558	32,000		Total	21,708	102,92	32,000		Total	20,405	11,595	32,000		Total	21,527	10,473	32,000	
Precision	0.7502	0.6025			Precision	0.7632	0.6817			Precision	0.787	0.7109			Precision	0.8332	0.7382			Precision	0.8344	0.8221		
F1 score	0.7742	0.5653			F1 score	0.8039	0.6093			F1 score	0.8181	0.6582			F1 score	0.8408	0.7267			F1 score	0.8688	0.7604		
Accuracy	0.7028				Accuracy	0.7389				Accuracy	0.7626				Accuracy	0.7988				Accuracy	0.8304			

0.4

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	17,514	3,227	20,741	0.8444	Ans=0	17,807	2,872	20,679	0.8611	Ans=0	17,912	2,776	20,688	0.8658	Ans=0	18,601	2,219	20,820	0.8934	Ans=0	18,891	1,920	20,811	0.9077
Ans=1	6,419	4,840	11,259	0.4299	Ans=1	4982	6,339	11,321	0.5599	Ans=1	3,943	7,369	11,312	0.6514	Ans=1	3,913	7,267	11,180	0.65	Ans=1	3,310	7,879	11,189	0.7042
Total	23,933	8,067	32,000		Total	22,789	9,211	32,000		Total	21,855	10,145	32000		Total	22,514	9,486	32,000		Total	22,201	9,799	32,000	
Precision	0.7318	0.6			Precision	0.7814	0.6882			Precision	0.8196	0.7264			Precision	0.8262	0.7661			Precision	0.8509	0.8041		
F1 score	0.7841	0.5009			F1 score	0.8193	0.6175			F1 score	0.8421	0.6869			F1 score	0.8585	0.7033			F1 score	0.8784	0.7508		
Accuracy	0.6986				Accuracy	0.7546				Accuracy	0.79				Accuracy	0.8084				Accuracy	0.8366			

0.5

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	18,617	2,531	21,148	0.8803	Ans=0	18,507	2,607	21,114	0.8765	Ans=0	18,985	2,276	21,261	0.8929	Ans=0	19,629	1,789	21,418	0.9165	Ans=0	19,767	1,841	21,608	0.9148
Ans=1	6,609	4,243	10,852	0.391	Ans=1	5,077	5,809	10,886	0.5336	Ans=1	4,365	6,374	10,739	0.5935	Ans=1	3,784	6,798	10,582	0.6424	Ans=1	3,125	7,267	10392	0.6993
Total	25,226	6,774	32,000		Total	23,584	8,416	32,000		Total	23,350	8,650	32,000		Total	23,413	8,587	32,000		Total	22,892	9,108	32,000	
Precision	0.738	0.6264			Precision	0.7847	0.6902			Precision	0.8131	0.7369			Precision	0.8384	0.7917			Precision	0.8635	0.7979		
F1 score	0.8029	0.4814			F1 score	0.8281	0.6019			F1 score	0.8511	0.6575			F1 score	0.8757	0.7093			F1 score	0.8884	0.7453		
Accuracy	0.7144				Accuracy	0.7599				Accuracy	0.7925				Accuracy	0.8258				Accuracy	0.8448			

0.6

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	19,768	1,766	21,534	0.918	Ans=0	18,949	2,655	21,604	0.8771	Ans=0	19,315	2,492	21,807	0.8857	Ans=0	20,302	1,724	22,026	0.9217	Ans=0	20,958	1,359	22,317	0.9391
Ans=1	7,375	3,091	10,466	0.2953	Ans=1	5,315	5,081	10,396	0.4887	Ans=1	4,724	5,469	10,193	0.5365	Ans=1	3,703	6,271	9,974	0.6287	Ans=1	3,621	6,062	9,683	0.626
Total	27,143	4,857	32,000		Total	24,264	7,736	32,000		Total	24,039	7,961	32,000		Total	24,005	7,995	32,000		Total	24,579	7,421	32,000	
Precision	0.7283	0.6364			Precision	0.781	0.6568			Precision	0.8035	0.687			Precision	0.8457	0.7844			Precision	0.8527	0.8169		
F1 score	0.8122	0.4034			F1 score	0.8262	0.5604			F1 score	0.8426	0.6025			F1 score	0.8821	0.698			F1 score	0.8938	0.7088		
Accuracy	0.7143				Accuracy	0.7509				Accuracy	0.7745				Accuracy	0.8304				Accuracy	0.8444			

0.7

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	20,678	1,346	22,024	0.9389	Ans=0	19,789	2,391	22,180	0.8922	Ans=0	20,711	1,693	22,404	0.9244	Ans=0	20,217	2,448	22,665	0.892	Ans=0	21,223	1,864	23,087	0.9193
Ans=1	7,311	2,665	9,976	0.2671	Ans=1	5,177	4,643	9,820	0.4728	Ans=1	4,619	4,977	9,596	0.5187	Ans=1	3,503	5,832	9,335	0.6247	Ans=1	3,156	5,757	8,913	0.6459
Total	27,989	4,011	32,000		Total	24,966	7,034	32,000		Total	25,330	6,670	32,000		Total	23,720	8,280	32,000		Total	24,379	7,621	32,000	
Precision	0.7388	0.6644			Precision	0.7926	0.6601			Precision	0.8176	0.7462			Precision	0.8523	0.7043			Precision	0.8705	0.7554		
F1 score	0.8269	0.3811			F1 score	0.8395	0.551			F1 score	0.8678	0.612			F1 score	0.8717	0.6622			F1 score	0.8942	0.6964		
Accuracy	0.7295				Accuracy	0.7635				Accuracy	0.8027				Accuracy	0.814				Accuracy	0.8431			

0.8

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	21,310	1,364	22,674	0.9398	Ans=0	21,072	1,890	22,962	0.9177	Ans=0	21,369	1,938	23,307	0.9168	Ans=0	22,630	1,524	24,154	0.9369	Ans=0	22,630	1,524	24,154	0.9369
Ans=1	7,072	2,254	9,326	0.2417	Ans=1	5,666	3,372	9,038	0.3731	Ans=1	4,480	4,213	8,693	0.4846	Ans=1	3,442	4,404	7,846	0.5613	Ans=1	3,442	4,404	7,846	0.5613
Total	28,382	3,618	32,000		Total	26,738	5,262	32,000		Total	25,849	6,151	32,000		Total	26,072	5,928	32,000		Total	26,072	5,928	32,000	
Precision	0.7508	0.623			Precision	0.7881	0.6408			Precision	0.8267	0.6849			Precision	0.868	0.7429			Precision	0.868	0.7429		
F1 score	0.8348	0.3483			F1 score	0.848	0.4716			F1 score	0.8694	0.5676			F1 score	0.9011	0.6395			F1 score	0.9011	0.6395		
Accuracy	0.7364				Accuracy	0.7639				Accuracy	0.7994				Accuracy	0.8448				Accuracy	0.8448			

0.9

32 x 32	Pred=0	Pred=1	Total	Recall	64 x 64	Pred=0	Pred=1	Total	Recall	128 x 128	Pred=0	Pred=1	Total	Recall	256 x 256	Pred=0	Pred=1	Total	Recall	512 x 512	Pred=0	Pred=1	Total	Recall
Ans=0	22,047	1,491	23,538	0.9367	Ans=0	22,950	1,345	24,295	0.9446	Ans=0	23,183	1,728	24,911	0.9306	Ans=0	24,022	1,513	25,535	0.9407	Ans=0	24,469	1,798	26,267	0.9315
Ans=1	6,538	1,924	8,462	0.2274	Ans=1	5,848	1,857	7,705	0.241	Ans=1	4,820	2,269	7,,089	0.3201	Ans=1	4,415	2,050	6,465	0.3171	Ans=1	3,449	2,284	5,733	0.3984
Total	28,585	3,415	32,000		Total	28,798	3,202	32,000		Total	28,003	3,997	32000		Total	28,437	3,563	32,000		Total	27,918	4,082	32,000	
Precision	0.7713	0.5634			Precision	0.7969	0.58			Precision	0.8279	0.5677			Precision	0.8447	0.5754			Precision	0.8765	0.5595		
F1 score	0.846	0.324			F1 score	0.8645	0.3405			F1 score	0.8763	0.4093			F1 score	0.8902	0.4089			F1 score	0.9032	0.4654		
Accuracy	0.7491				Accuracy	0.7752				Accuracy	0.7954				Accuracy	0.8147				Accuracy	0.836			

Supplementary Table 3

Pred = “Prediction” : 0 = “unstained” , 1 = “stained”
Ans = “Answer” : 0 = “unstained” , 1 = “stained”

Input block size (px)

Target block size: 32 x 32 (px)

32 x 32

	Pred=0	Pred=1	Total	Recall
Ans=0	17,680	2,705	20,385	0.8673
Ans=1	6,259	5,356	11,615	0.4611
Total	23,939	8,061	32,000	
Precision	0.7385	0.6644		
F1 score	0.7978	0.5444		
Accuracy	0.7199			

64 x 64

	Pred=0	Pred=1	Total	Recall
Ans=0	17,991	2,394	20,385	0.8826
Ans=1	5,533	6,082	11,615	0.5236
Total	23,524	8,476	32,000	
Precision	0.7648	0.7176		
F1 score	0.8195	0.6054		
Accuracy	0.7523			

128 x 128

	Pred=0	Pred=1	Total	Recall
Ans=0	17,781	2,604	20,385	0.8723
Ans=1	4,420	7,195	11,615	0.6195
Total	22,201	9,799	32,000	
Precision	0.8009	0.7343		
F1 score	0.8351	0.672		
Accuracy	0.7805			

256 x 256

	Pred=0	Pred=1	Total	Recall
Ans=0	17,888	2,497	20,385	0.8775
Ans=1	3,656	7,959	11,615	0.6852
Total	21,544	10,456	32,000	
Precision	0.8303	0.7612		
F1 score	0.8533	0.7212		
Accuracy	0.8077			

512 x 512

	Pred=0	Pred=1	Total	Recall
Ans=0	18,397	1,988	20,385	0.9025
Ans=1	3,173	8,442	11,615	0.7268
Total	21,570	10,430	32,000	
Precision	0.8529	0.8094		
F1 score	0.877	0.7659		
Accuracy	0.8387			

Target block size: 128 x 128 (px)

128 x 128

	Pred=0	Pred=1	Total	Recall
Ans=0	18,409	2,852	21,261	0.8659
Ans=1	3,388	7,351	10,739	0.6845
Total	21,797	10,203	32,000	
Precision	0.8446	0.7205		
F1 score	0.8551	0.702		
Accuracy	0.805			

256 x 256

	Pred=0	Pred=1	Total	Recall
Ans=0	18,187	1,873	20,060	0.9066
Ans=1	3,946	7,994	11,940	0.6695
Total	22,133	9,867	32,000	
Precision	0.8217	0.8102		
F1 score	0.8621	0.7332		
Accuracy	0.8182			

512 x 512

	Pred=0	Pred=1	Total	Recall
Ans=0	18,356	1,704	20,060	0.9151
Ans=1	3,645	8,295	11,940	0.6947
Total	22,001	9,999	32,000	
Precision	0.8343	0.8296		
F1 score	0.8728	0.7562		
Accuracy	0.8328			

Supplementary Table 4

Pred = “Prediction” : 0 = “unstained” , 1 = “stained”
Ans= “Answer” : 0 = “unstained” , 1 = “stained”

A

Automatically binarized

Small network

	Pred=0	Pred=1	Total	Recall
Ans=0	18,043	1,783	19,826	0.9101
Ans=1	2,724	9,450	12,174	0.7762
Total	20,767	11,233	32,000	
Precision	0.8688	0.8413		
F1 score	0.889	0.8075		
Accuracy	0.8592			

Large network

	Pred=0	Pred=1	Total	Recall
Ans=0	18,291	1,535	19,826	0.9226
Ans=1	1,574	10,600	12,174	0.8707
total	19,865	12,135	32,000	
Precision	0.9208	0.8735		
F1 score	0.9217	0.8721		
Accuracy	0.9028			

Rebinarized

Small network

	Pred=0	Pred=1	Total	Recall
Ans=0	23,556	974	24,530	0.9603
Ans=1	1,354	6,116	7,470	0.8187
Total	24,910	7,090	32,000	
Precision	0.9456	0.8626		
F1 score	0.9529	0.8401		
Accuracy	0.9273			

Large network

	Pred=0	Pred=1	Total	Recall
Ans=0	23,927	603	24,530	0.9754
Ans=1	622	6,848	7,470	0.9167
Total	24,549	7,451	32,000	
Precision	0.9747	0.9191		
F1 score	0.975	0.9179		
Accuracy	0.9617			

B

Small network Automatically binarized / Large network Rebinarized

Fold 1

Small network, Automatically binarized

	Pred=0	Pred=1	Total	Recall
Ans=0	17,764	12,928	30,692	0.5788
Ans=1	488	8,820	9,308	0.9476
Total	18,252	21,748	40,000	
Precision	0.9733	0.4056		
F1 score	0.7259	0.568		
Accuracy	0.6646			

Large network, Rebinarized

	Pred=0	Pred=1	Total	Recall
Ans=0	34,853	2,740	37,593	0.9271
Ans=1	972	1,435	2,407	0.5962
Total	35,825	4,175	40,000	
Precision	0.9729	0.3437		
F1 score	0.9494	0.436		
Accuracy	0.9072			

Fold 2

Small network, Automatically binarized

	Pred=0	Pred=1	Total	Recall
Ans=0	14,204	3,980	18,184	0.7811
Ans=1	3,679	18,137	21,816	0.8314
Total	17,883	22,117	40,000	
Precision	0.7943	0.82		
F1 score	0.7876	0.8257		
Accuracy	0.8085			

Large network, Rebinarized

	Pred=0	Pred=1	Total	Recall
Ans=0	18,096	1,505	19,601	0.9232
Ans=1	3,984	16,415	20,399	0.8047
Total	22,080	17,920	40,000	
Precision	0.8196	0.916		
F1 score	0.8683	0.8568		
Accuracy	0.8628			

Fold 3

Small network, Automatically binarized

	Pred=0	Pred=1	Total	Recall
Ans=0	20,037	1,934	21,971	0.912
Ans=1	6,803	11,226	18,029	0.6227
Total	26,840	13,160	40,000	
Precision	0.7465	0.853		
F1 score	0.821	0.7199		
Accuracy	0.7816			

Large network, Rebinarized

	Pred=0	Pred=1	Total	Recall
Ans=0	28,847	749	29,596	0.9747
Ans=1	2,245	8,159	10,404	0.7842
Total	31,092	8,908	40,000	
Precision	0.9278	0.9159		
F1 score	0.9507	0.845		
Accuracy	0.9252			

Fold 4

Small network, Automatically binarized

	Pred=0	Pred=1	Total	Recall
Ans=0	23,713	5,601	29,314	0.8089
Ans=1	3,116	7,570	10,686	0.7084
Total	26,829	13,171	40,000	
Precision	0.8839	0.5747		
F1 score	0.8447	0.6346		
Accuracy	0.7821			

Large network, Rebinarized

	Pred=0	Pred=1	Total	Recall
Ans=0	34,920	952	35,872	0.9735
Ans=1	734	3,394	4,128	0.8222
Total	35,654	4,346	40,000	
Precision	0.9794	0.7809		
F1 score	0.9764	0.801		
Accuracy	0.9578			