Classifier of iPSCs Using Machine Learning and Deep Learning-Based Approaches

Abstract

Induced pluripotent stem cells (iPSCs) hold great potential for regenerative medicine and disease modeling. Accurate classification of iPSCs and their differentiated morphology is crucial for advancing these applications. In this project, we developed a machine learning-based classifier to differentiate iPSC with different morphology using a publicly available dataset. We applied a deep learning model, CNN and optimized it using RMSprop(Root Mean Square Propagation).By combine the conventional machine learning model, including SVM(support vector machine), PCA(Principal component analysis) and LDA(Linear Discriminant Analysis), the classifier demonstrated high accuracy.

Backgroung

Induced pluripotent stem cells (iPSCs) are a type of stem cell derived by reprogramming somatic cells, granting them the ability to differentiate into various cell types. This remarkable characteristic makes iPSCs highly promising for applications in regenerative medicine, disease modeling, and drug screening. Since iPSCs can be directly derived from a patient's own tissues and reprogrammed, they are particularly valuable for personalized medicine. iPSCs offer the potential for tissue replacement or repair, enable the study of disease mechanisms, and provide platforms for the development of new pharmaceuticals.

However, accurate identification and classification of iPSCs are critical for advancing these applications. The iPSCs can be classify by its morphology. The accurate classification of these different cell types not only enhances the safety and efficacy of stem cell therapies but also improves the precision of disease models.

The manual classification and observation of iPSC morphology are time-consuming and subjective processes. Therefore, the aim of this project is developing a machine learning-based classifier to automatically distinguish these different cell types, which could significantly improve classification accuracy and accelerate research progress.

Method

Dataset

In this study, we used a dataset of morphologically annotated single-cell images of human induced pluripotent stem cells (iPSCs), which were used for deep learning applications. The dataset contains images of iPSCs in various cellular states, captured using high-content automated imaging. The cells were stained with several markers, including DAPI and bright-field imaging, to capture different biological properties and cellular features. These

images were collected from 2020 to 2022, with a total of 5962 images in iPSC quality control dataset and 1312 images in iPSC morphology dataset.

In the iPSC morphology dataset, each image is 150*150*7channel. The dataset consists of seven channels: 647 (alpha-Tubulin), Brightfield, DAPI, 488 (OCT4), 555 (GSK3-beta), NuclearSegmentation, and CellSegmentation. The 647 channel highlights microtubules, the Brightfield channel provides general cell structure, DAPI stains cell nuclei, and the 488 and 555 channels capture the expression of pluripotency marker OCT4 and GSK3-beta, respectively(Fig. 1).

The dataset includes five distinct labels representing different cellular morphologies of iPSCs: **Big**, indicating larger and spread-out cells; **Long**, for elongated cells; **Mitotic**, marking cells in the process of division; **RAR-treated**, for cells treated with retinoic acid receptor agonists; and **Round**, for cells with a rounded shape. These labels were assigned to single-cell sub-images based on their morphological characteristics, facilitating the classification of iPSC states for machine learning analysis.

The dataset is publicly available under a Creative Commons Attribution-ShareAlike 4.0 International License and can be accessed through the ETH Zurich repository (https://doi.org/10.3929/ethz-b-000581447). It provides a valuable resource for training machine learning models, especially for tasks such as single-cell phenotyping and the classification of iPSC differentiation.

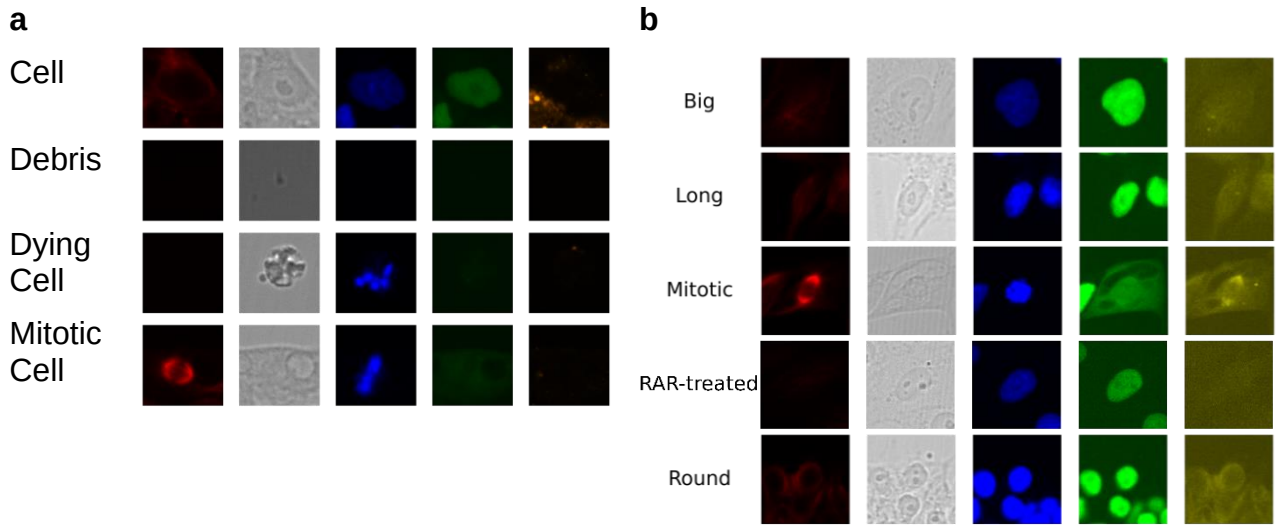

Fig. 1 Representative images of each label in the dataset (a) iPSC quality control dataset (b) iPSC morphology dataset

Data argumentation

The models were tested in two configurations: with and without data augmentation. For the data augmentation approach, three transformations were applied to the images: vertical flipping, horizontal flipping, and both.

Model

In this project, we implemented four different model architectures for classifying iPSC morphologies:

CNN + Dense: A convolutional neural network (CNN) is used to automatically extract features from the images, followed by a fully connected dense layer for classification. This model leverages the spatial features captured by the CNN for decision-making.

CNN + SVM: In this approach, CNN is used for feature extraction, and the features are then passed to a Support Vector Machine (SVM) for classification. The SVM is utilized to improve the decision boundary between classes by finding the optimal hyperplane in the feature space.

CNN + PCA + SVM: This model combines CNN for feature extraction with Principal Component Analysis (PCA) for dimensionality reduction, followed by SVM for classification. PCA helps reduce the computational complexity by selecting the most relevant features, making the SVM more efficient.

CNN + LDA + SVM: Similar to the previous model, CNN is used for feature extraction, followed by Linear Discriminant Analysis (LDA) for dimensionality reduction. LDA is employed to maximize class separability before passing the features to the SVM for classification.

Each model aims to leverage the strengths of CNN for feature extraction while combining it with different classification techniques to optimize performance.(Fig. 2)

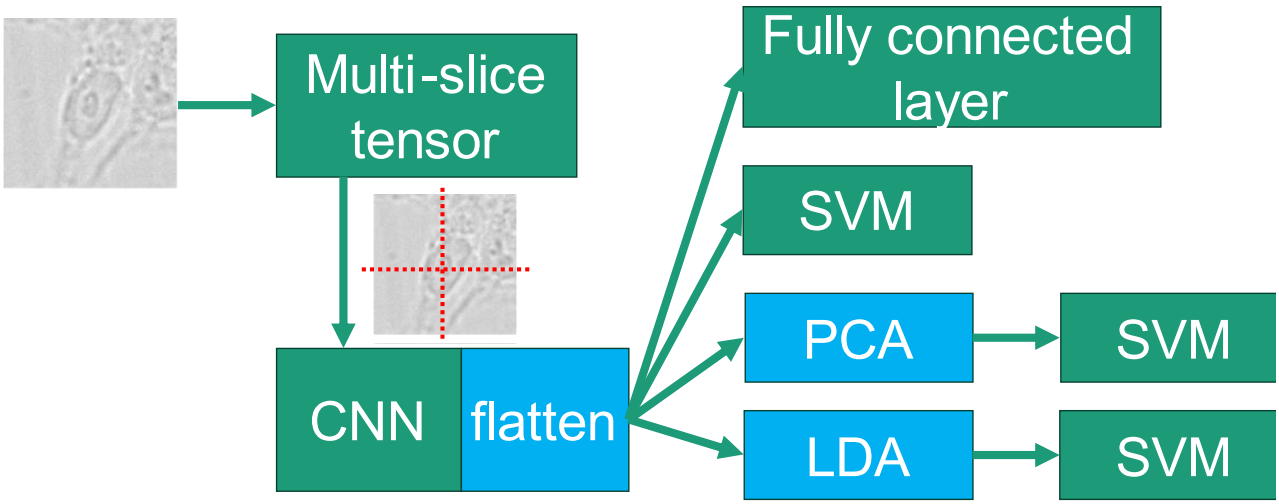

Fig. 2 Relationship of the four models used in the study: CNN + Dense, CNN + SVM, CNN + PCA + SVM, and CNN + LDA + SVM.

Hyperparameters

The dense layers consist of two 1024-unit layers, each followed by a dropout of 0.2 to prevent overfitting. The CNN model uses a 3x3 kernel for each single tensor, with filters progressively increasing from 32 to 64 to 128 layers. The model employs SAME padding and utilizes the ReLU activation function for all layers. In the PCA-based model, the Radial Basis

Function (RBF) kernel is used with 64 components (n=64), and a gamma value of 1e-5 to effectively capture the underlying structure of the data. For the LDA model, we select 4 components (n=4) to maximize class separability. Finally, the SVM classifier is set with a polynomial kernel of degree 3, a regularization parameter (C) of 100, coef0 set to 1, and the one-vs-one decision function shape to handle multi-class classification. These hyperparameters are tuned to optimize the classification performance of the models.

Evaluation

The performance of the models was evaluated using accuracy as the primary metric. Additionally, training history, including accuracy and loss curves, was monitored during the training process to assess model convergence. To ensure the robustness of the models, experiments were repeated with multiple random seeds, and statistical comparisons were performed to check for significant differences in performance across runs.

Result

accuracy

For the models without data augmentation, the accuracy results show notable variation across the different configurations, which demonstrate an average accuracy ranging from 0.7270 to 0.8653, with the highest accuracy observed at 0.8653 for the CNN+SVM model.

In contrast, the models with data augmentation generally show more stable and higher accuracy values, with averages ranging from 0.8037 to 0.892. These group also show higher accuracy than one in without data augmentation group. These models exhibit improved stability and performance, suggesting that data augmentation plays a significant role in enhancing the model's generalization capability and reducing overfitting. It get the highest accuracy of 0.892 in CNN+SVM model

The overall trend highlights that data augmentation is crucial for improving the model’s performance, particularly for tasks requiring high accuracy and generalization, as seen in the consistent improvement of accuracy with the inclusion of augmentation techniques.

The trend within the group shows that CNN+SVM > CNN+PCA+SVM > CNN+LDA+SVM, with all showing significant differences. Regarding the fully connected layer, it shows the lowest accuracy in the group without data augmentation. However, it does not show a significant difference compared to CNN+SVM, which has the highest accuracy.

Training history

From the training history, it is evident that the model without data augmentation (Fig. 4b) suffered from poor validation loss performance, which failed to decrease throughout the training process. In contrast, when data augmentation was applied, the validation loss showed a significant and consistent decline(Fig. 4d). Furthermore, the accuracy comparison revealed that in the model without data augmentation, the validation accuracy fluctuated

erratically, indicating instability(Fig. 4a). However, with data augmentation, the validation accuracy demonstrated a more stable and improved trend, highlighting the importance of data augmentation in reducing overfitting and enhancing model generalization. (Fig. 4c)

feature extraction of CNN

The CNN effectively extracted meaningful features from the input images. By leveraging convolutional layers with 3×3 kernels and progressively increasing filters (32→64 →128)(Fig. 5a), the network captured spatial hierarchies and morphological details of iPSCs. These features were critical for subsequent classification, as demonstrated by the improved performance of combined models, particularly those utilizing PCA and LDA for dimensionality reduction. The results highlight the ability of the CNN to distill high-dimensional data into discriminative representations for accurate classification. All 128 highest layer results shown in Fig. 5b.

dimension reduction

Dimension reduction results show that the 64,000-dimensional vector output from CNN, when visualized using UMAP, indicates that features are being captured effectively. In PCA, data points are separated by color, highlighting the distinction between different classes. LDA, on the other hand, exhibits a stronger tendency to form clusters, further enhancing class separation.(Fig. 6)

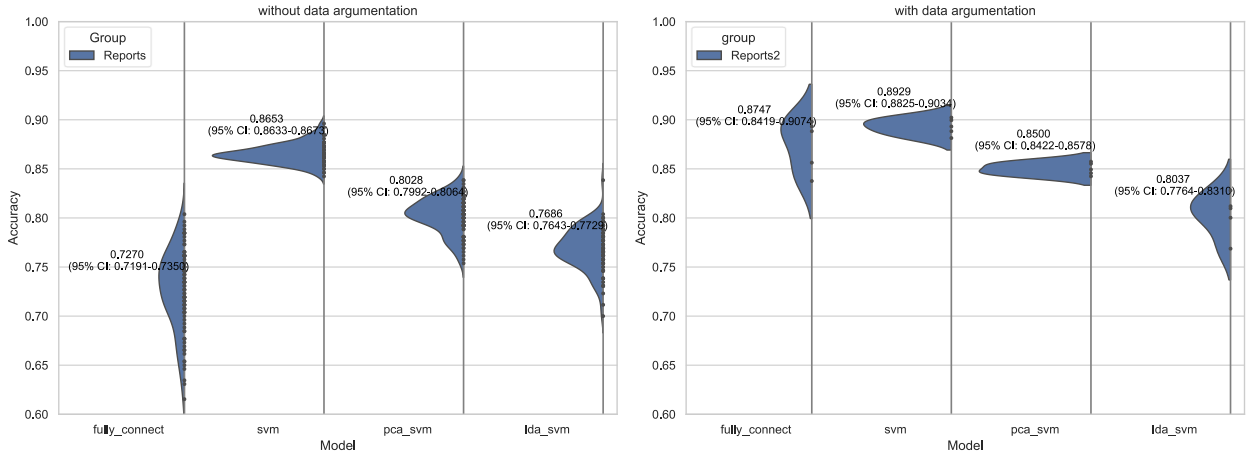

Fig. 2 compares model performance with and without data augmentation. CNN+SVM outperforms the other models

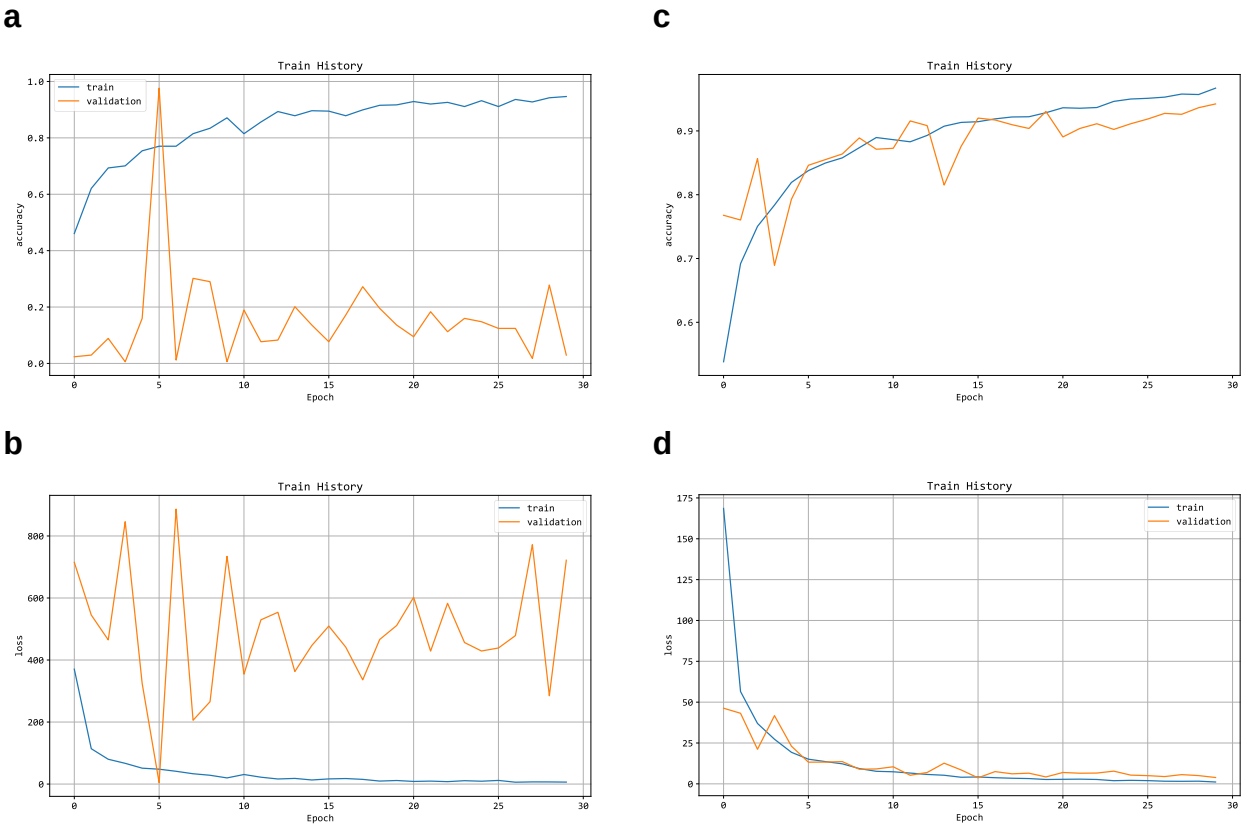

Fig. 4: Training history comparison of models with and without data augmentation. (**a**) Validation accuracy without data augmentation, showing instability. (**b**) Validation loss without data augmentation, failing to decrease. (**c**) Validation accuracy with data augmentation, showing stability. (**d**) Validation loss with data augmentation, decreasing consistently.

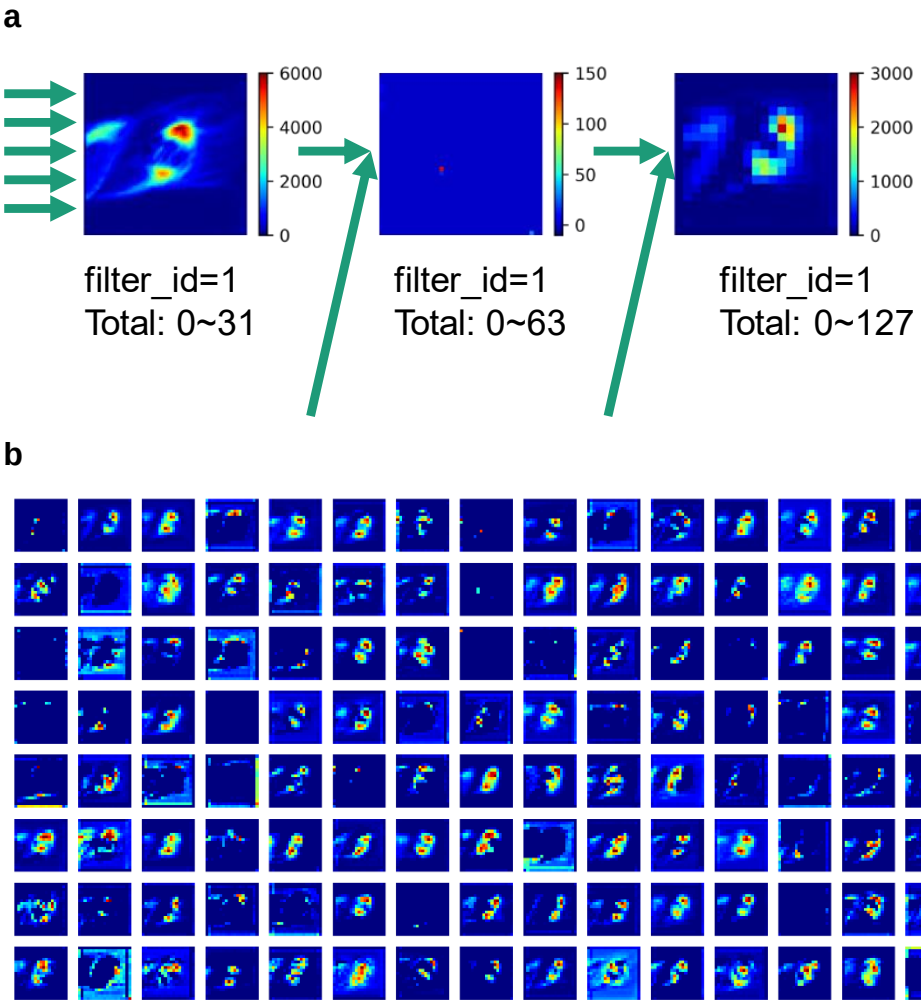

Fig. 5 (a) Visualization of the convolutional layers with progressively increasing filters (32→64→128) that capture spatial hierarchies and morphological details of iPSCs. (b) Features extracted from the final 128-filter layer, showcasing the high-level discriminative representations used for classification.

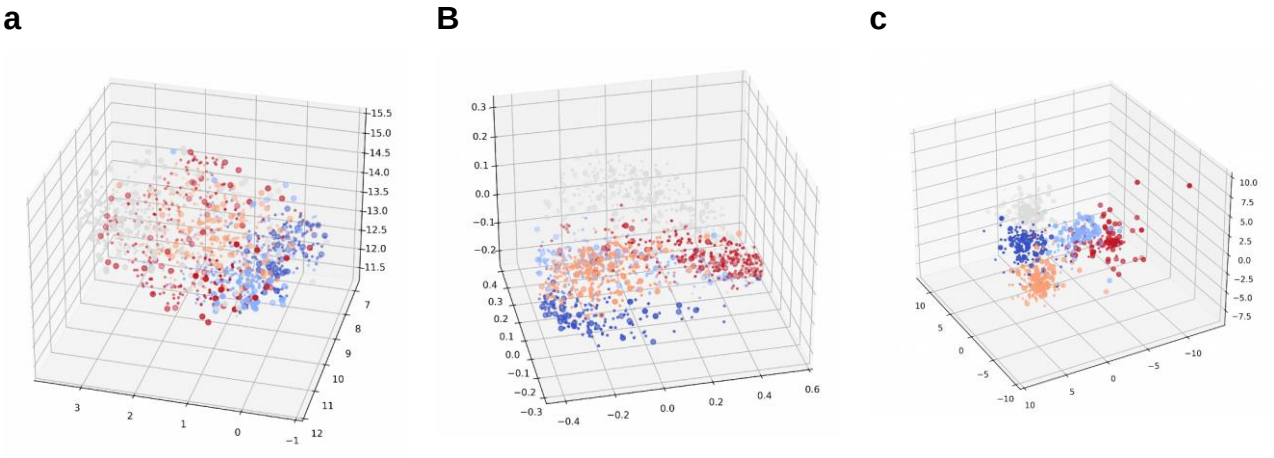

Fig. 6: (a) UMAP visualization of the CNN output, showing captured features. (b) PCA projection, where different colors represent distinct data clusters. (c) LDA plot, highlighting stronger clustering and improved class separation.

Discussion

Optimizer

The commonly used Adam optimizer did not perform well in this project, failing to achieve satisfactory convergence or accuracy. Instead, RMSprop with a learning rate of 1e-5 was found to be more effective for this task. RMSprop, with its adaptive learning rate and ability to handle the varying gradients of different parameters, provided more stable and efficient training. This highlights the importance of selecting the right optimizer based on the task at hand and underscores the need for careful tuning of hyperparameters to achieve optimal performance.

Overfitting and data argumentation

Without data augmentation, the model exhibited significant overfitting, as evidenced by the validation loss remaining very high and failing to decrease during training. This suggests that the model struggled to generalize well to unseen data. However, when data augmentation was applied, including vertical, horizontal flips and both, the validation loss showed a clear reduction, indicating improved generalization. The use of data augmentation helped the model better handle variations in the data, mitigating overfitting and resulting in more robust performance on the validation set.

PCA and LDA does not improve the accuracy

The inclusion of PCA and LDA did not improve the overall accuracy compared to other methods. Generally, the performance ranking observed was CNN+SVM > CNN+Dense > CNN+PCA+SVM > CNN+LDA+SVM.

This suggests that dimensionality reduction using PCA or LDA may not effectively capture key features for classification in this dataset, potentially due to the loss of important high-dimensional information.

Conclusion

In this study, we explored various model architectures for classifying iPSC morphologies, demonstrating the effectiveness of combining CNN with different classification techniques. Data augmentation was crucial in improving model generalization and reducing overfitting. Additionally, the use of RMSprop as the optimizer with a learning rate of 1e-5 enhanced training stability, leading to better performance. These findings highlight the importance of optimizing model components for accurate and robust classification in biological applications.

1. *Download Cell Data: Images, Genomics, & Features*. (2018). ALLEN CELL EXPLORER. https://www.allencell.org/data-downloading.html#DownloadFeatureData

2. Lien, C.-Y., Chen, T.-T., Tsai, E.-T., Hsiao, Y., Lee, N., Gao, C.-E., Yang, Y., Chen, S., Yarmishyn, A. A., Hwang, D., Chou, S., Chu, W.-C., Chiou, S., & Chien, Y. (2023). Recognizing the Differentiation Degree of Human Induced Pluripotent Stem Cell-Derived Retinal Pigment Epithelium Cells Using Machine Learning and Deep Learning-Based Approaches. *Cells*, *12*(2), 211–211. https://doi.org/10.3390/cells12020211

3. Pfaendler, R. (2022). Morphologically annotated single-cell images of human induced pluripotent stem cells for deep learning. Ethz.ch. https://doi.org/10.3929/ethz-b-000581447