Predicting Induced Pluripotent Stem Cell Derived Cardiomyocytes Differentiation Outcomes

Angela Zhang
angelaz@stanford.edu

1 Introduction: Project Motivation and Problem Statement

Induced pluripotent stem cell derived cardiomyocytes (iPSC-CMs) have revolutionized cardiovascular research. iPSC-CMs are patient specific pluripotent stem cells (iPSCs), cells capable of becoming any cell type in the body, that have been differentiated into heart cells [1]. In the past decade iPSC-CMs have been used as an in-vitro platform to study **cardiovascular disease mechanisms and drug induced cardiotoxicity**[2]. However, iPSC-CMs harbor a bottleneck that limits their wider clinical translation: **creating the large amounts of iPSC-CMs needed for disease modeling and drug screening is laborious, time consuming, imprecise (subject to batch to batch variation), and expensive.** This is because differentiating iPSCs into heart cells frequently results in a heterogenous population of cells that can result in high, medium, low, or no yield of the desired heart cells. Concurrently, differentiating iPSCs into heart cells is a 3 week process and there is no way to know prospectively the differentiation yield of heart cells. Thus, time and resources are wasted in this process, limiting the scalability of iPSC-CMs. Recently it was suggested that morphology cues obtained through non-invasive easy to obtain phase contrast images early in the differentiation process can be used to predict differentiation outcomes ie yield of heart cells [1-2]. Early prediction of differentiation outcomes can save time and resources. **Thus, the goal of this proposal is to develop an image classifier to predict differentiation heart cell yield from an input of phase contrast images during early stage differentiation. (Figure. 1)**

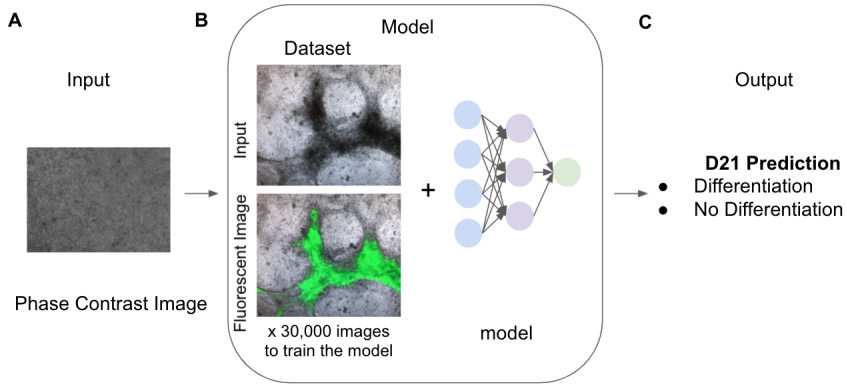


Figure 1: **An overview of the proposed project.** The input is a phase contrast image taken during an early stage (i.e. D7) differentiation, while the desired output is a classification of whether or not the image will result in cardiac cells at the end of the 3 week process (i.e. D21). The fluorescent image shows the heart cells in green. The fluorescent image provides per pixel labels and will be used to label the outcome.

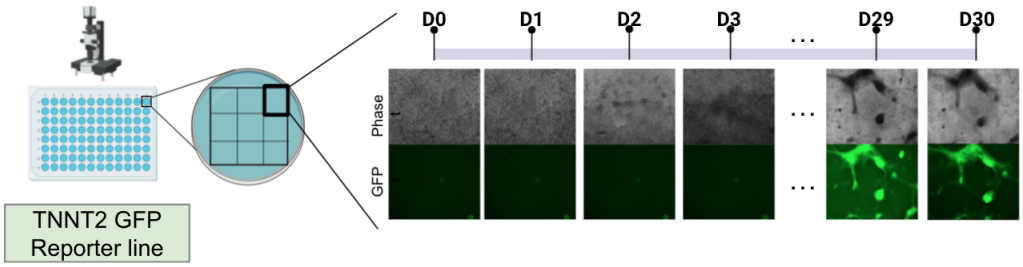
CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Dataset

This project uses a **dataset of 30,000 images** that I have previously created. The dataset consists of phase contrast images of iPSC-CMs and their corresponding fluorescent images (Figure 1). The flourescent images provide automated and standardized labels for the differentiation outcomes. The flourescence in an image indicates the presence of a heart cell (green in Figure 1). Thus, the flourescent image and the presence of fluorescence will be used to label the differentiation outcomes–whether there are heart cells are not. The dataset was created by capturing 3,000 phase contrast and fluorescent images each day of the 21 day differentiation of iPSCs into heart cells. Each image is **2MB and 1224x904 pixels and the dataset is balanced** (Figure 2). The dataset as seen in Figure 2 demonstrates morphological changes during the 21 day differentiation process. Importantly, later during the differentiation the morphology between differentiated and non-differentiated cells becomes more pronounced, which ideally can be captured by the classifier.

Dataset Preprocessing and Augmentation The images were collected from 9 different locations of the well, and thus have different illuminations based on location. I pre-processed the data according to location in the well (ie. I preprocessed all corner images together, all center images together). I performed global intensity normalization for the images. The per-image pixel intensity distributions were constrained to have a fixed mean and standard deviation. To fit pre-trained computer vision models, images were reformatted to 224x224 and 229x229 for Inception.

A



B

Characterization of Dataset			
Total Number of Images	Size of image	Dimensions	% of images that differentiated
30,000	2.1MB Total: 330 GB	1224 x 904 pixels	Class Balanced 44% differentiation 56% no differentiation

Figure 2: **A schematic of the how the dataset was created and characteristics of the dataset.** Phase contrast images (model inputs) and fluorescent images (input labels) were taken during each day of the 21 day differentiation process.

3 Approach

Baseline Standards and Problem Premise. Current standards can determine differentiation outcomes at D21 with accuracy of 90 percent [1]. The goal is then to (1) build a classifier to meet current standards and (2) to build a classifier to predict differentiation outcomes at an earlier time point. To do this, I will train 9 models–one model for every two days during differentiation (Figure 3). For example, data from D3 and 4 will be used to train the D3-4 model and so on. This allows me to create a model to match current standards (1) and determine the earliest time point that differentiation can be predicted (2).

Model Development. I utilized **transfer learning** and pre-trained Convolutional Neural Networks to develop the 9 models. Using the dataset, I retrained the last classification layer. I explored **multiple pre-trained CNN architectures**, including ResNet, VGG, InceptionV3, and DenseNet, but saw non-significant change in performance. ResNet was found to result in more consistent performance, and therefore it was selected as the primary model architecture for the 9 models (Appendix 1). I also experimented with **various optimization functions**, including SGD, SGD with

momentum, ADAgrad and Adam. Unsurprisingly SGD performed the worst consistently across all 9 models, however suprisingly SGD momentum was found to outperform Adam. To further improve performance, **hyperparameter tuning/search** was performed. The primary hyperparameters that were tuned were : **learning rate, momentum, beta, and number of training epochs**. As taught in CS 230, I performed **random hyperparameter search but on a log scale**. Learning rate and momentum were found to be the most significant hyperparameters (or the hyperparameters I managed to tune correctly...). Initially the model performance would stagnate quickly or oscillate–decreasing the learning rate improved performance while delivering an optimal learning curve. Finally, there was severe overfitting initially, however, through **augmenting the dataset** through horizontal flip, vertical flip, and random rotation, overfitting decreased significantly. In fact, data augmentation, particularly random rotation, resulted in the largest increase in performance (greater than any hyperparameter tuning or change in model architecture or optimizer). **To summarize, 9 models were developed using pre-trained ResNet architecture and SGD with momentum optimizer.**

4 Results

Model Performance. Model performance of the 9 trained models were evaluated by examining the train/validate loss and accuracy curves; precision and recall; F1 score; and confusion matrix. Figure 3 gives an overview of model results.

Project Goal 1. By evaluating the accuracy and F1 results for the D18-21 model, we see that we were able to meet current baseline standards of around 90 percent prediction of heart cell yield on D21.

Project Goal 2. I next examined the accuracy and F1 curves to discover the earliest day that differentiation outcomes can be predicted with approximate baseline performance. We see that the D7-8 model yields an accuracy of 0.83 and and F1 score of 0.85, which is close to the baseline performance. Furthermore, we see a significant drop in performance in models trained on data from days earlier than D7-8. This result is consistent with biological mechanisms. It has been previously shown that cell fate is committed to heart cells around D6-7.

Analysis A confusion matrix of the D7-8 model demonstrates that the model has high true positive and precision, low false positive, and slightly high false negative rate. This is ideal as in practice it is better to precisely identify cells that will become heart cells, and it is more costly (reagents and time) to have high false positive rate.

Conclusion. The developed D18-21 model meets existing standards while the D7-8 model is able to predict differentiation outcomes up to 2 weeks earlier than current standards with only a moderate drop in performance.

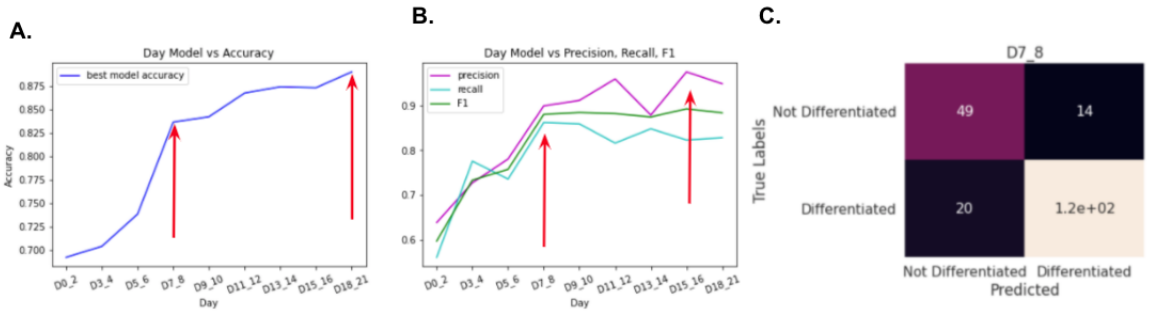


Figure 3: **Evaluation of model reveals high performance and high precision.** 9 CNN classifiers were developed and trained: one model for every two days (ie D3-4, D5-6 etc.). (A) and (B) display the test accuracy, precision, recall and F1 for each of the 9 trained models. Both (A) and (B) demonstrate that model performance at D18-21 is able to meet the performance of current gold standards. When answering the question of when is the earliest day that differentiation can be predicted reliably, we see that both (A) and (B) suggest that D7-8 may be the earliest time point. After D7-8 (ie D0-6) model performance drops significantly. (C) Confusion matrix of D7-8 model demonstrates that the model has high true positive and precision.

Model Interpretation: Saliency Map. To interpret model performance, saliency map analysis was performed for several test images (Figure 4). When comparing the saliency map with the fluorescent image, which demonstrates where the heart cells are (white), we see that the highest intensity regions of the saliency map (yellow ring) correlate with the highest intensity regions of the fluorescent map (yellow ring). This trend is seen across multiple examples. This offers some relief that the model is not training on an artifact, and perhaps suggests that the model has learned biologically relevant features. Note that the fluorescent models were not used to train the model, and therefore there is no risk of data leakage.

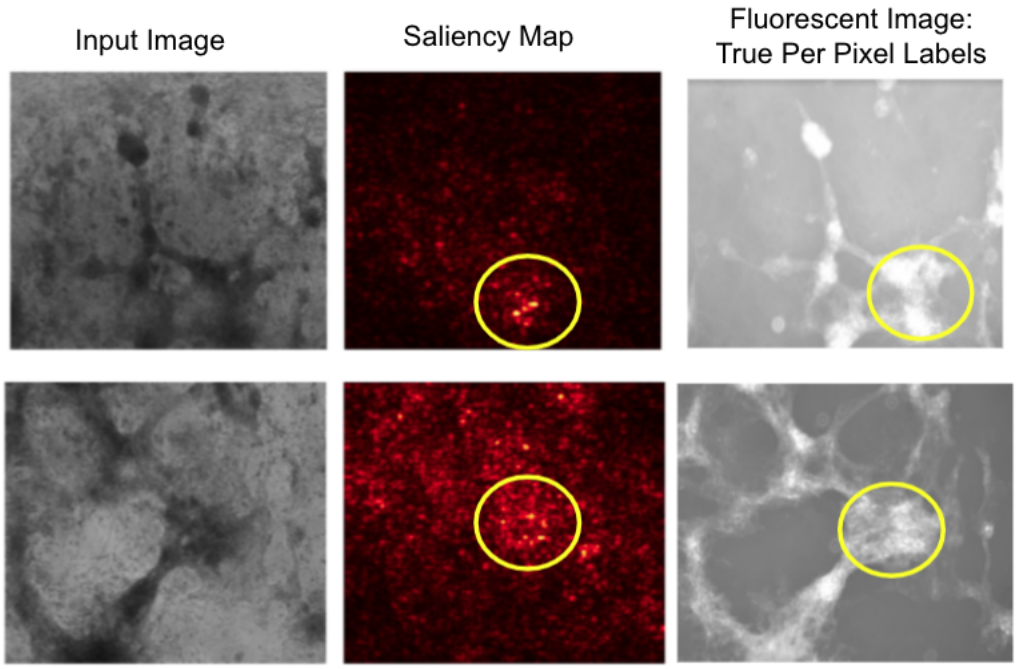


Figure 4: **Saliency map reveals model detects biologically relevant regions.** High intensity regions of the saliency map (yellow ring) correlate with the highest intensity regions of the flourescent map (yellow ring). This suggests that the model has learned biologically relevant features. Note that the fluorescent models were not used to train the model, and therefore there is no risk of data leakage.

Model Generalizability. One significant obstacle that current deep learning models face is the ability to generalize or maintain high performance on datasets that were not used to train the model. To evaluate the generalizability of the developed models, I evaluated the model's performance on an additional test dataset. The train dataset was created from iPSCs derived from one individuals, while the test dataset includes data generated from 3 additional individuals. Figure 4a demonstrates that iPSCs generated from different patients may have variations in morphology. Thus, evaluating the model on an external test dataset will allow me to determine (i) model generalizability and (ii) whether the learned pattern is a general biological pattern or an individual specific artifact. Figure 4b shows accuracy and F1 score of the train-test and external test dataset. Model performance decreases, which is to be expected; however, the general trends of the 9 models are maintained–ie D7-8 remains an inflection point. Figure 4c provides a closer breakdown of model performance. We find that while true negative rate remains high, the true positive rate decreases. Given the importance of precision at the proposed task, this reveals limitations for using the model for additional individual lines.

5 Discussion and Future Work.

To summarize, I developed 9 models using a pre-trained ResNet model and was able to match existing performance. Furthermore, one of the developed models D7-8 is able to match existing performance but predict outcomes up to 2 weeks prior. I then utilized saliency map to determine that the model

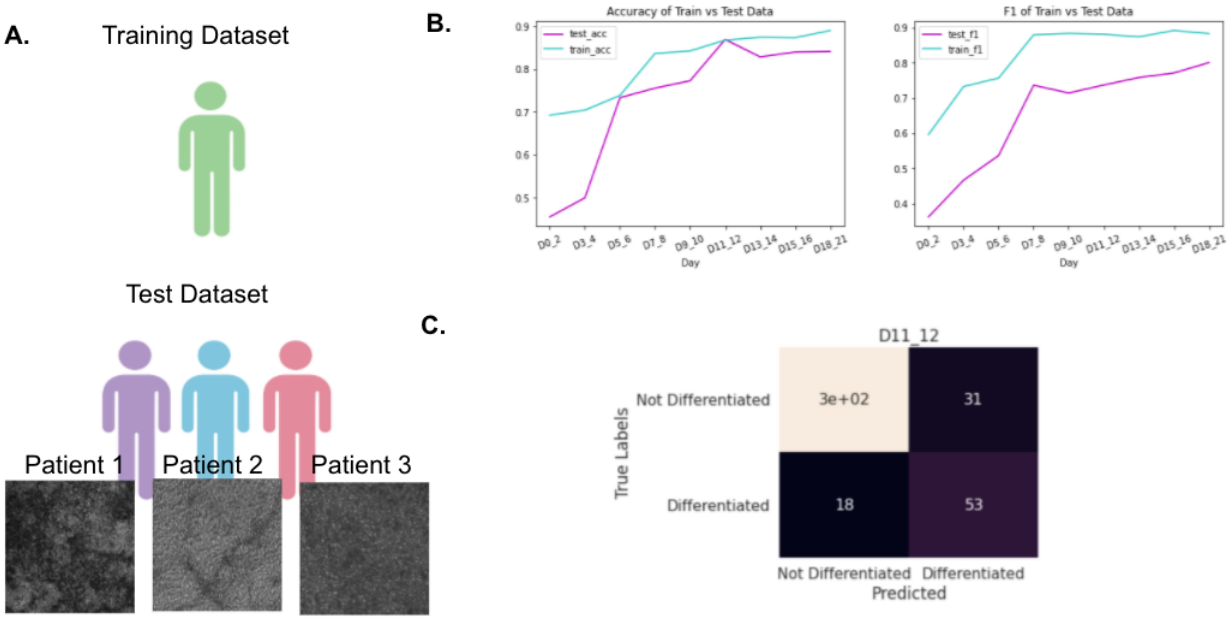


Figure 5: **Model demonstrates potential to generalize.** Using data from 3 additional patients that are different from the patient data used to train and develop the model, model generalizability was evaluated. (A) demonstrates that patient specific induced pluripotent stem cells can be morphologically different. (B) and (C) compare the accuracy and F1 score between the original train dataset and the test dataset. We find that model performance deteriorates, which is to be expected, but generally maintains the same patterns–ie Day 7-8 is still an inflection point. (C) Confusion matrix gives insight into deterioration of performance–precision has decreased.

regions of interest correlate with biological read outs. Finally, I evaluated the generalizability of the model using an external test dataset.

One limitation of the project is the use of saliency map for model interpretation. Work done by Adebayo et al. in "Sanity Checks for Saliency Maps" reveals that most saliency maps produce similar results despite models with inputs of random noise or random weights [8]. Given this finding, I should utilize alternative interpretation mechanisms such as using influence functions to identify the data most responsible for a given prediction [9].

Future Work will involve improving model performance and generalizability. Prior reports demonstrated that creating a segmentation model with an addtional classification model can deliver greater performance than classification models [6]. Furthermore, it was demonstrated that these models can combat model deterioriation when evaluating model generalizability.

References

[1] Chen, H., Zhang, A. and Wu, J.C., 2018. Harnessing cell pluripotency for cardiovascular regenerative medicine. Nature biomedical engineering, 2(6), pp.392-398.

[2] Christiansen, E.M., Yang, S.J., Ando, D.M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O'Neil, A., Shah, K., Lee, A.K. and Goyal, P., 2018. In silico labeling: predicting fluorescent labels in unlabeled images. Cell, 173(3), pp.792-803.

[3] Ounkomol, Chawin, et al. "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy." Nature methods 15.11 (2018): 917-920.

[4] Buggenthin, Felix, et al. "Prospective identification of hematopoietic lineage choice by deep learning." Nature methods 14.4 (2017): 403-406.

[5] 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(7639), pp.115-118.

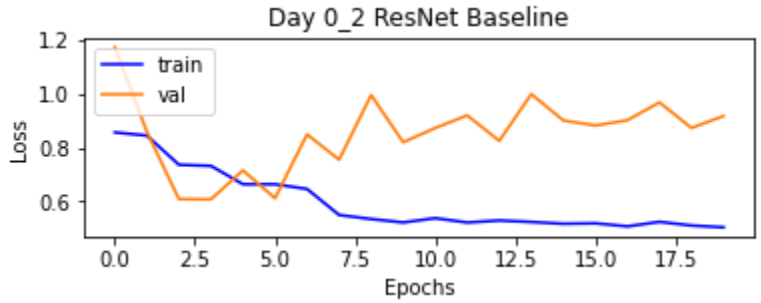
[6] De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D. and van den Driessche, G., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine, 24(9), pp.1342-1350.

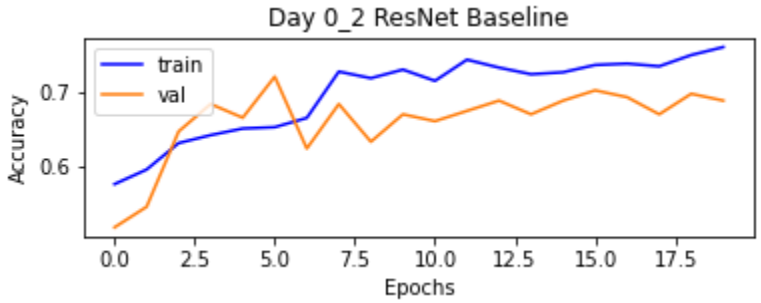
[7] Yi, X., Walia, E. and Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Medical image analysis, 58, p.101552.

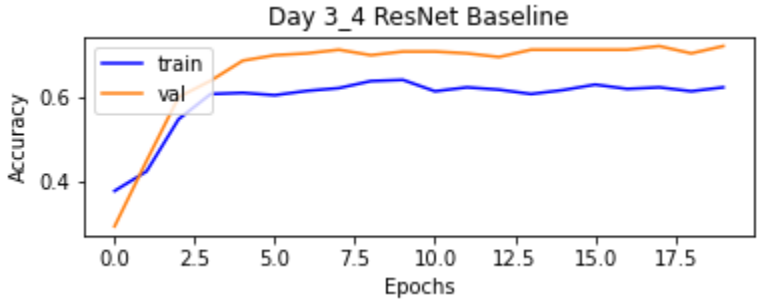
[8] Adebayo, Julius, et al. "Sanity checks for saliency maps." arXiv preprint arXiv:1810.03292 (2018).

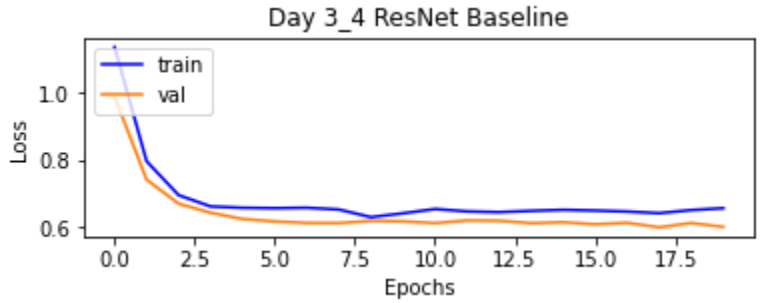
[9] Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." International Conference on Machine Learning. PMLR, 2017.

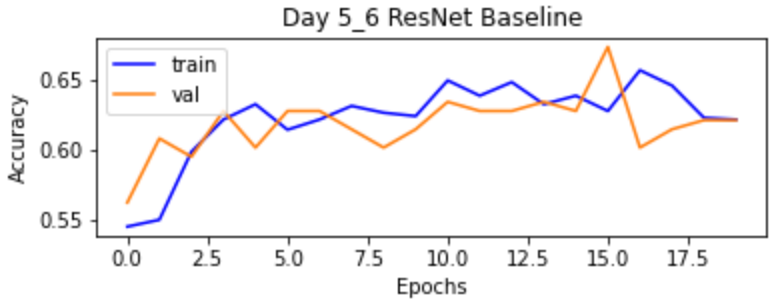
Appendix 1.

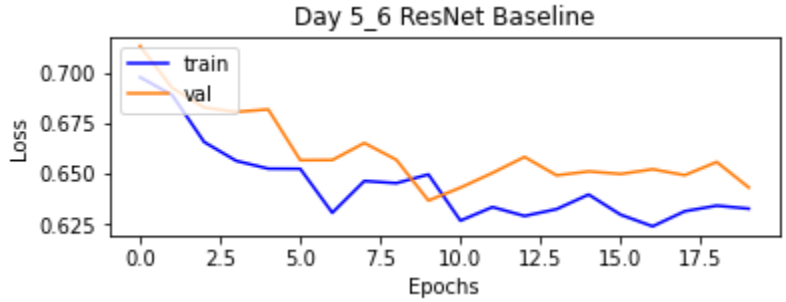

Day 0_2 ResNet Baseline
train
val
Loss
Epochs

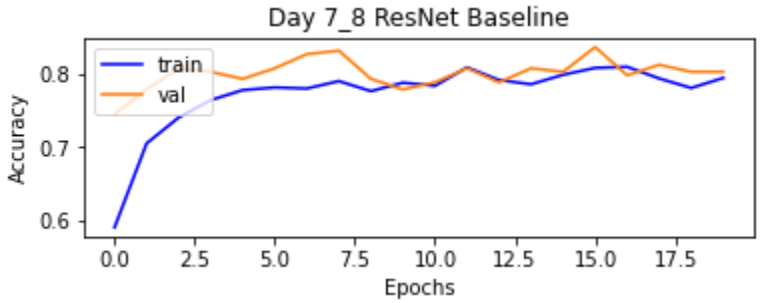

Day 0_2 ResNet Baseline
train
val
Accuracy
Epochs

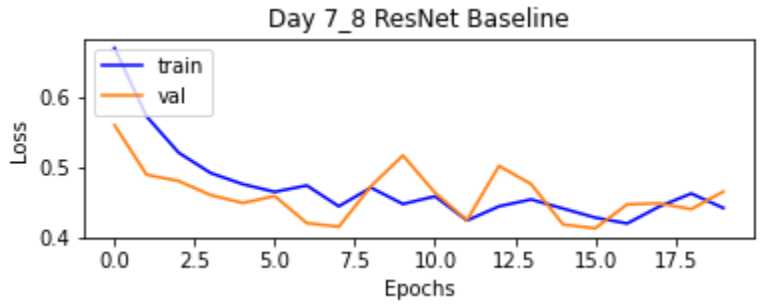

Day 3_4 ResNet Baseline
train
val
Accuracy
Epochs

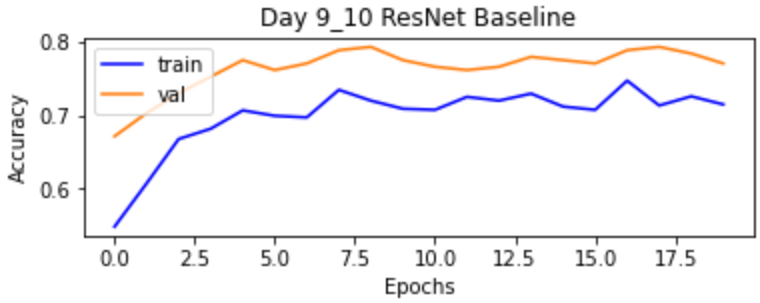

Day 3_4 ResNet Baseline
train
val
Loss
Epochs

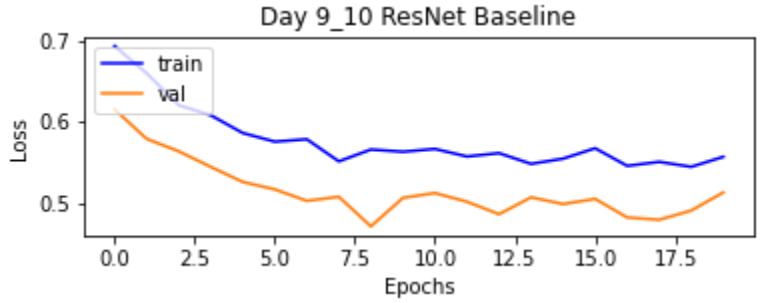

Day 5_6 ResNet Baseline
train
val
Accuracy
0.65
0.60
0.55
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs

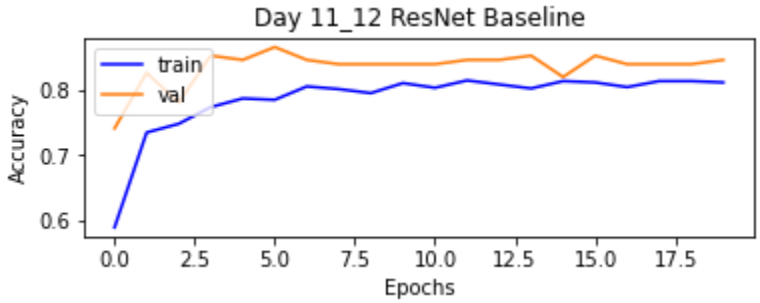

Day 5_6 ResNet Baseline
train
val
Loss
0.700
0.675
0.650
0.625
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs

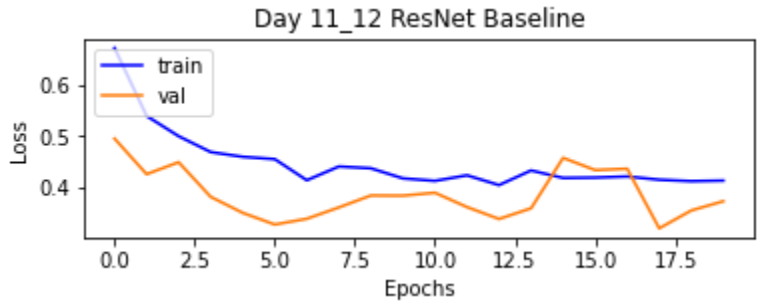

Day 7_8 ResNet Baseline
train
val
Accuracy
0.8
0.7
0.6
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs

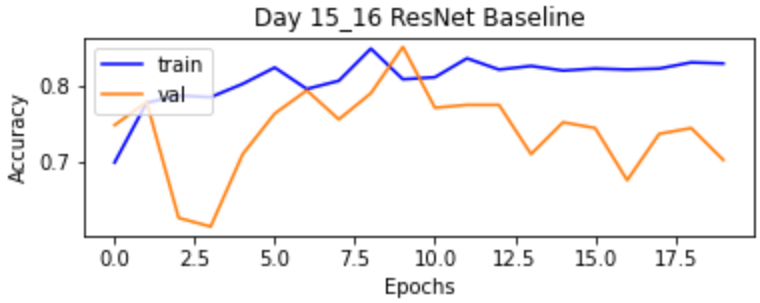

Day 7_8 ResNet Baseline
train
val
Loss
0.6
0.5
0.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs

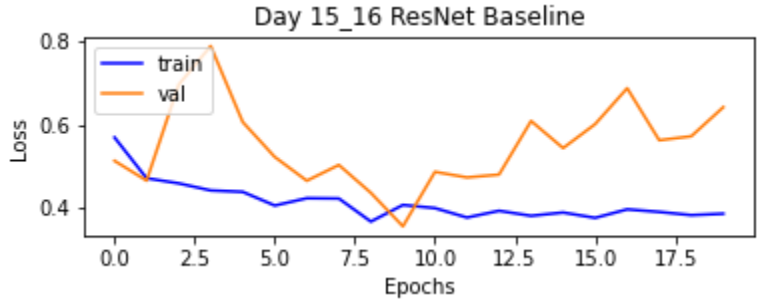

Day 9_10 ResNet Baseline
train
val
Accuracy
0.8
0.7
0.6
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs

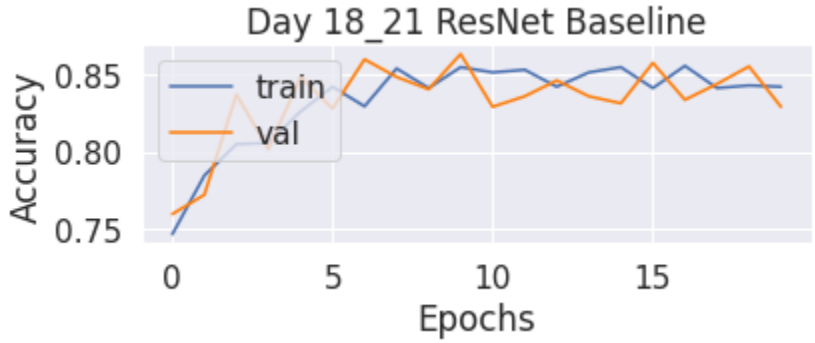

Day 9_10 ResNet Baseline
train
val
Loss
0.7
0.6
0.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs

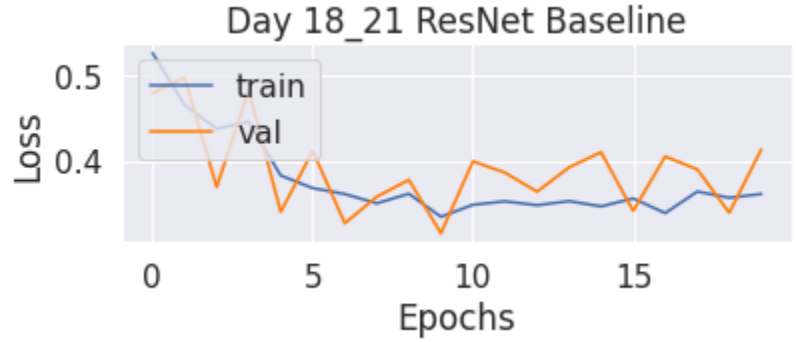

Day 11_12 ResNet Baseline
train
val
Accuracy
0.8
0.7
0.6
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs


Day 11_12 ResNet Baseline
train
val
Loss
0.6
0.5
0.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs


Day 15_16 ResNet Baseline
train
val
Accuracy
0.8
0.7
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs


Day 15_16 ResNet Baseline
train
val
Loss
0.8
0.6
0.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs


Day 18_21 ResNet Baseline
train
val
Accuracy
0.85
0.80
0.75
0
5
10
15
Epochs


Day 18_21 ResNet Baseline
train
val
Loss
0.5
0.4
0
5
10
15
Epochs