Project 2: Predicting Drug Treatments from Cell Images

Dhruv Jani (dj2688)

STAT 5243 - Applied Data Science
Prof. Bianca Dumitrascu

Abstract

This project explores the use of machine learning and image processing techniques to predict drug treatments from cell images. Using a subset of a publicly available dataset of three million cell images, we analyzed morphological changes in cells treated with various drugs. Features were extracted from microscopy images using Cellpose for segmentation, and a Random Forest classifier was trained to predict drug treatments. Preliminary results demonstrate the ability to distinguish treated and control (DMSO) groups with high accuracy. These findings highlight the potential of data-driven methods for analyzing cellular phenotypes.

1 Introduction

Understanding how drugs affect cellular morphology is a key problem in cell biology. In this project, we analyze images of cells treated with chemical and genetic perturbations to predict the specific drug treatment based on morphological features. The dataset used in this project was generated using fluorescent dyes to label cellular components such as the nucleus, cytoskeleton, and mitochondria, as described in [1].

The primary goals of this project were:

- To preprocess and extract meaningful features from cellular images.
- To train and evaluate a classifier for drug treatment prediction.
- To interpret the results and identify potential patterns in cellular responses.

2 Methods

2.1 Dataset

We utilized a curated subset(2867 images) of the dataset provided in [1]. The dataset includes metadata with information about the perturbations applied to each image.

2.2 Image Segmentation and Feature Extraction

Cellpose [2] was used for cell segmentation. Features such as cell area, perimeter, and eccentricity were extracted for each segmented cell. The feature extraction pipeline was implemented in Python and processed images in batches for computational efficiency.

2.3 Model Training and Evaluation

A Random Forest classifier was used to predict the drug treatments. The steps included:

- Splitting the dataset into training and testing sets.
- Encoding labels for drug treatments.
- Evaluating the classifier using metrics such as accuracy, precision, and recall.

2.4 Code Availability

The complete codebase is available at: Project 2.

3 Results

3.1 Feature Extraction

Cellpose successfully segmented the majority of images, and morphological features were aggregated for downstream analysis.

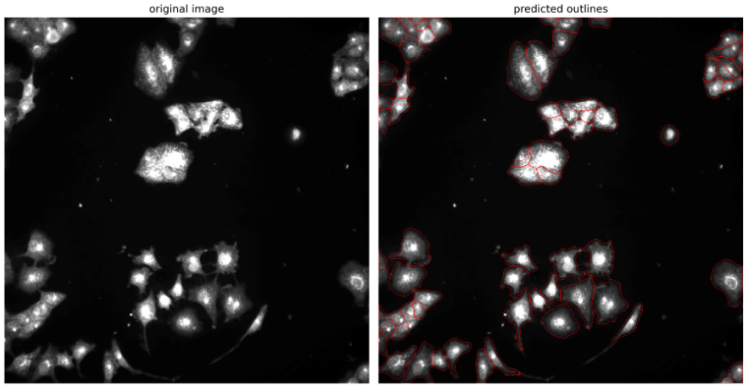

Figure 1: Cell segmentation using Cellpose: Original Images & Predicted Outline.

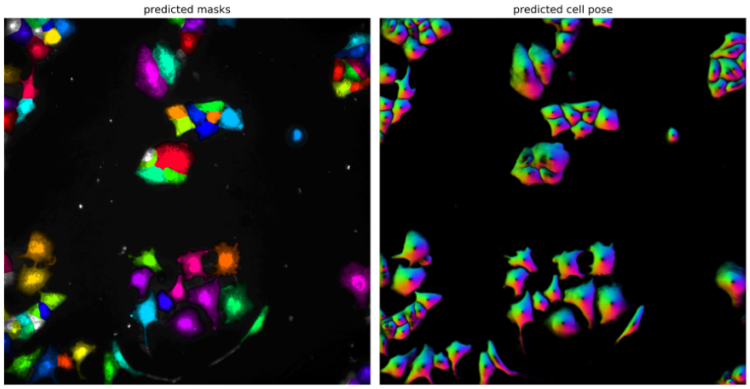

Figure 2: Cell segmentation: Predicted Masks & Predicted Cellpose

3.2 Classification Performance

The Random Forest model achieved a classification accuracy of **70%** on the test set.

3.3 Model Output

The Random Forest model predicts the treatment for any input cell image after cellpose segmentation, feature extraction & analysis.

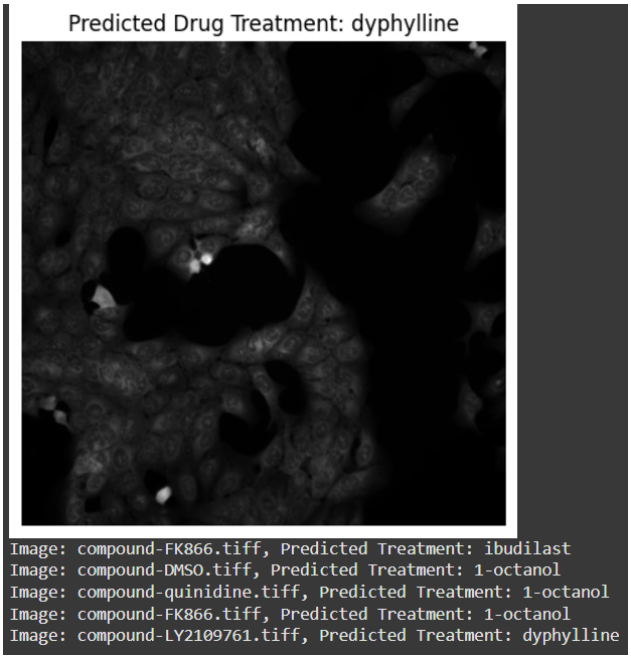

Figure 3: Model output for drug treatment predictions.

4 Conclusion

In this project, we demonstrated the feasibility of predicting drug treatments from cell images using a machine learning pipeline. Our approach leveraged Cellpose for segmentation and morphological feature extraction, followed by Random Forest classification. Future work could explore deep learning methods for feature extraction and prediction, as well as the biological interpretation of the results.

References

[1] Mark-Anthony Bray, Shantanu Singh, Han Han, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 13:935–940, 2016.

[2] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: A generalist algorithm for cellular segmentation. *bioRxiv*, 2020. Preprint.