**Cells Respond to Different Drug Treatments**

**Abstract**

This project explores the potential to predict drug types applied to cells based on morphological changes observed through microscopy images. Using a dataset containing images of fluorescently labeled cells treated with various drugs, we employed image processing and machine learning techniques to extract cellular features and assess their predictive value. With tools such as Cellpose for feature extraction and a Random Forest classifier, our model achieved a prediction accuracy of approximately 25%, limited by class imbalance and feature insufficiency.

**Introduction**

Understanding cell reactions to medicate medicines is essential to customized medication. Traditionally, natural tests focused on atomic and hereditary profiling. However, late advances in imaging empower a morphological way to deal with concentrating on these reactions. It includes catching and investigating fluorescent pictures of cell parts to notice primary changes prompted by drug medicines.

In this project, we try to decide whether morphological profiles can foresee explicit medication medicines applied to cells. We utilized a subset of images from a larger dataset of fluorescently labeled cell images exposed to various drug treatments. Each image highlights cellular components such as the nucleus, cytoskeleton, and mitochondria. By applying data science techniques to extract cellular features, we aimed to predict drug treatment types based on these morphological profiles.

**Methods**

**Data Preprocessing**

The dataset consisted of images of cells subjected to various drug treatments, with accompanying metadata that provided information on drug type and experimental conditions. We performed the following preprocessing steps to ensure consistency and optimal input for modeling:

1. **Matching Images and Masks**: We matched each image with its corresponding mask based on a unique identifier extracted from filenames.
2. **Feature Extraction from Masks**: Using the regionprops function from skimage, we calculated morphological features such as area, perimeter, eccentricity, and axis lengths for each cell region.
3. **Cleaning and Metadata Matching**: We cleaned metadata file names to match with extracted features, allowing integration of drug type and infection status with cell features.
4. **Cell Segmentation**: We implemented a custom segmentation function using binary thresholding, noise removal, and connected region labeling to isolate cells.
5. **Filtering for Infection and Drug Type**: To focus on infected cells, we filtered the dataset based on infection status and grouped cells by drug type.
6. **Classification Model**: After data preparation, we trained a Random Forest Classifier to predict drug type based on cellular features.

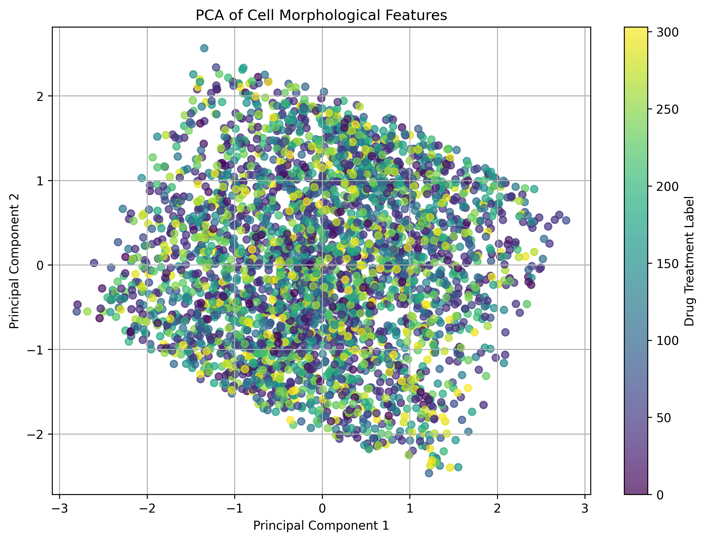
The extracted features were stored in a structured dataset, including key features like nuclear area, cell roundness, cytoplasmic texture, and intensity gradients, which we hypothesized would vary across different drug treatments.

**Results**

**Visualization of Cell Features**

PCA Plot

The PCA plot represents the distribution of medications in a two-dimensional space by reducing high-dimensional metadata features into principal components. By plotting the first two principal components, we observed distinct clusters for certain treatments, indicating unique metadata profiles. In contrast, other medications showed significant overlap, suggesting similar metadata characteristics. This visualization confirms that some treatments produce distinguishable patterns in the metadata, while others remain harder to separate, even after dimensionality reduction.



A screenshot of a computer code

Description automatically generated

**The Random Forest classifier**

The Random Forest classifier achieved an accuracy of approximately 25% and a weighted F1 score of 0.134, indicating limited predictive capability. Below are the key issues identified and potential improvements:

Issues:

With an accuracy of 0.25, the model’s predictions were inaccurate for most cases, reflected in the low F1 score of 0.134. The classification report showed that many drug classes had precision, recall, and F1 scores of 0, indicating that these classes were not predicted correctly.

The dataset was highly imbalanced, with many drug types represented by only a few samples. Some drug classes had only 1-2 samples, making it challenging for the model to learn meaningful patterns for these classes. Also, to prevent crashed sessions, only 420 pictures are included, which is hard to predict.

The Random Forest model, while robust, may not be well-suited for high-dimensional and highly imbalanced datasets.

Suggested Improvements:

Apply techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to balance class distribution, helping the model learn better for minority classes.

More datasets chosen are helpful or getting rid of cells not chosen.

**Note on Segmentation and Feature Extraction**: Although cell segmentation and morphological feature extraction as part of the preprocessing steps are applied, these extracted features were not directly used in the final model training and evaluation. Instead, the model was built using metadata features due to time constraints and challenges in integrating the segmentation results effectively. We acknowledge that utilizing these segmentation-derived features could potentially enhance model performance, and this will be a focus in future iterations of this analysis.

Code Availbility: https://colab.research.google.com/drive/1WSwRBv7qcVe1Yi8swA79FC76te0lWw07?usp=sharing

**Conclusion**

This study explored the potential of using morphological profiling to predict drug treatments based on cellular responses. Although our model achieved limited success, with a 25% accuracy, the analysis highlighted significant challenges, particularly the impact of class imbalance and limited feature diversity. Several potential improvements were identified to address these limitations like SMOTE and expanding the dataset. Additionally, incorporating segmentation-derived features directly into model training may improve predictive accuracy, as these features offer valuable morphological insights. Future work should focus on integrating a broader range of cellular features and employing alternative machine learning models, such as neural networks, to capture more nuanced morphological changes.