Predicting Drug Treatment Effects on Cell Morphology Using Machine Learning on Image-Based Features

Abstract

This study investigates the use of morphological features derived from cell images to predict drug treatments, focusing on distinguishing between DMSO-treated (control) cells and those treated with other compounds. Using Cellpose for cell segmentation and a pre-trained ResNet50-based Convolutional Neural Network (CNN) for feature extraction, we trained and evaluated two classification models: a Random Forest and a Support Vector Machine (SVM) with an RBF kernel. Both models achieved moderate accuracy, with the SVM model showing a slight advantage in classifying DMSO samples correctly. The UMAP visualization revealed substantial overlap between DMSO and non-DMSO samples, indicating that morphological differences between these treatments may be subtle. This overlap limited the models' ability to fully distinguish the two groups, suggesting that morphological features alone may not capture the nuanced cellular responses induced by various drug treatments.

Introduction

Understanding the cellular effects of various drug treatments is crucial in drug discovery and development, as it provides insights into how specific compounds alter cell morphology and behavior. Identifying these changes can help in predicting drug mechanisms, assessing potential therapeutic effects, and uncovering unwanted side effects. In real-world applications, such as high-throughput drug screening, it is often desirable to use image-based techniques to evaluate large numbers of compounds quickly. Cell morphology, captured through microscopy imaging, offers a promising way to achieve this goal, as morphological changes can often reflect underlying biological processes influenced by drug treatments. However, identifying subtle morphological differences induced by various treatments, especially between control (e.g., DMSO) and experimental compounds, remains a challenging task.

This project builds upon the foundation established by the study "Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations" published in Nature. The original study aimed to systematically investigate the morphological profiles of cells treated with a broad spectrum of drugs and genetic perturbations. By applying image segmentation and feature extraction methods to cell images, the study demonstrated that morphological features could reveal information about drug mechanism of action (MoA) and genetic changes. Following a similar approach, this project focuses on a subset of cell images treated with DMSO and other compounds. The goal is to explore whether morphological features derived from these images can be used to predict the type of treatment applied, with particular emphasis on distinguishing between DMSO-treated (control) and non-DMSO-treated cells.

Methods

Data Processing

The dataset we used in this experiment are 2867 median aggregated images over the first 3 channels of one of the six batches of CPJUMP1 experiments described in the paper. The data is consisted of cell images treated with various chemical compounds, including DMSO as a control treatment and several other drugs of interest.

To prepare the images for analysis, we applied Cellpose for segmentation using the cyto2 model. This segmentation step was essential for identifying individual cell boundaries within each image

and generating corresponding cell masks. For segmentation, we set the diameter to 100 pixels to match typical cell size and specified channel to be 0 for segmentation since the images only had one grayscale channel. These settings allowed the model to accurately capture the cell morphology. Additionally, Cellpose produced flow images alongside the masks, which represented directional information about cell morphology. Both the segmented masks and flow images served as the basis for subsequent feature extraction, allowing us to capture structural and morphological characteristics relevant to the classification task.

Following segmentation, features were extracted from the processed images to capture morphological details indicative of each treatment. A Convolutional Neural Network (CNN) based on the pre-trained ResNet50 model was utilized for this purpose. The CNN model was applied to the segmented images to extract high-dimensional feature vectors, which captured complex visual patterns associated with different drug treatments. This step resulted in feature representations for each image that formed the input data for classification. Since the number of rows of metadata is different from the number of dataset we have, we extract the key words from segmented image file names and the column called FileName_OrigRNA in metadata to align them for further use.

Analysis

The analysis objective was to classify each image as either DMSO-treated or non-DMSO-treated. Given the class imbalance, with DMSO samples being underrepresented, we addressed this issue by adjusting class weights during model training. This approach increased the model's sensitivity to the DMSO class by penalizing misclassifications of DMSO samples more heavily. Two classification models were used to explore this binary classification task: a Random Forest model and a Support Vector Machine (SVM) model.

The Random Forest model was chosen for its robustness and ability to handle high-dimensional data, making it suitable for complex feature sets derived from cell morphology. In parallel, the SVM model was employed with an RBF kernel, which allows for non-linear decision boundaries and can help with distinguishing subtle differences between DMSO and non-DMSO treatments.

To evaluate the performance of these models, we utilized accuracy scores and confusion matrices. Accuracy scores provided an overall measure of each model's effectiveness, while the confusion matrices offered insights into the distribution of correct and incorrect predictions across the DMSO and non-DMSO classes. By examining the confusion matrices, we identified any patterns of misclassification, particularly noting whether the models struggled to correctly identify DMSO samples. The comparison of accuracy and confusion matrices between the Random Forest and SVM models allowed us to assess which model was more effective for the classification task.

Code Availability

Cell segmentation code are available at https://github.com/Quinnie2959/GR5243Project2/blob/8c13205f31ee7df60deb62221579736929b0f759/Cellpose_cell_segmentation_2D_prediction_only.ipynb, while data processing and model testing code are available at: https://github.com/Quinnie2959/GR5243Project2/blob/8c13205f31ee7df60deb62221579736929b0f759/Project2.ipynb.

Results

The performance of the Random Forest and SVM models was assessed through accuracy scores and confusion matrices, which provided insight into each model's ability to correctly classify DMSO

and non-DMSO samples. The Random Forest model achieved an accuracy of 81.533%, while the SVM model yielded an accuracy of 63.240%. Although both models performed similarly in terms of overall accuracy, their confusion matrices reveal distinct patterns of misclassification.

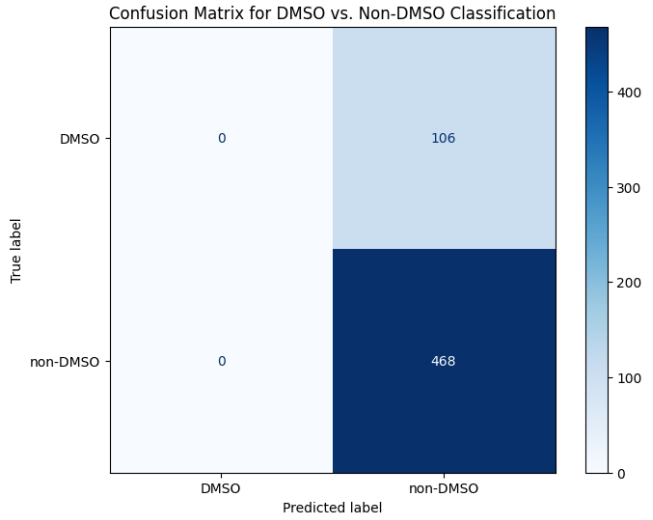

Figure 1. Confusion matrix for the Random Forest model

As shown in Figure 1, the confusion matrix for the Random Forest model demonstrates that it struggled to correctly identify DMSO samples. Specifically, the model frequently misclassified DMSO-treated cells as non-DMSO, resulting in a high false-negative rate for the DMSO class. This finding suggests that the Random Forest model had difficulty capturing morphological patterns unique to DMSO-treated cells, possibly due to the subtlety of morphological changes or class imbalance in the dataset.

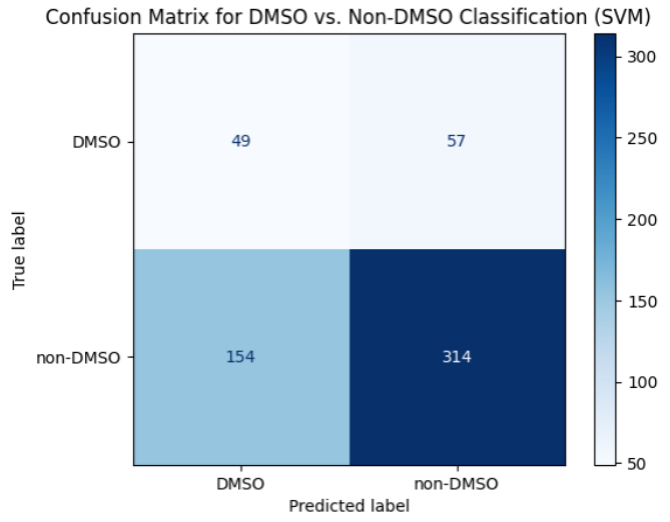

Figure 2. Confusion matrix for the SVM model

In contrast, the confusion matrix for the SVM model, as in Figure 2, shows a more balanced performance, with a greater number of correct classifications for both DMSO and non-DMSO samples. The SVM model demonstrated a slightly better sensitivity to DMSO samples, resulting in fewer false negatives compared to the Random Forest. This difference in performance may indicate that the SVM model was better suited for capturing non-linear relationships in the high-dimensional

feature space created by the CNN, which may contain complex morphological patterns relevant to distinguishing DMSO from non-DMSO treatments.

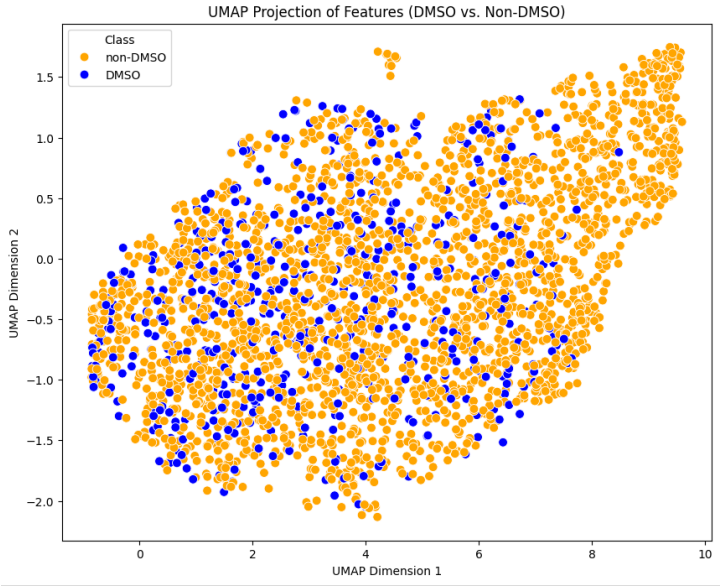

Figure 3. UMAP visualization of feature space

To further understand the separability of DMSO and non-DMSO samples based on morphological features, we visualized the feature space after dimensionality reduction using UMAP. Figure 3 illustrates the clustering of DMSO and non-DMSO samples in this reduced feature space. The DMSO samples were intermixed with non-DMSO samples, which indicates significant overlap in morphological patterns between the two classes. This overlap likely contributed to the models' misclassification patterns and underscores the subtlety of morphological differences between DMSO and non-DMSO treatments.

Conclusion

In summary, while both the Random Forest and SVM models achieved moderate accuracy in distinguishing DMSO from non-DMSO samples, the SVM model showed a slight advantage in identifying DMSO-treated cells. Our findings suggest that morphological features alone provide some information for distinguishing DMSO-treated cells, but the overlap between DMSO and non-DMSO samples in the feature space indicates substantial similarity in cell morphology between these groups. This overlap may account for the models' misclassifications, especially in distinguishing DMSO from certain non-DMSO treatments. These results underscore the challenge of using morphology alone to differentiate between subtle cellular responses, as the morphological differences induced by DMSO and other treatments may be minimal or difficult to capture with traditional imaging techniques.

While using a laptop with not enough RAM and unstable GPU access, it took a really long time to compute the data in the desired way and whole session was pretty easy to crash. If faster computation speed or better conditions we could obtain, there are several improvements that could be done. The CNN-based feature extraction was effective to an extent, but additional preprocessing steps or advanced feature extraction techniques, such as fine-tuning the CNN on drug-specific data, could potentially capture more subtle morphological changes induced by different treatments. Exploring multi-channel imaging may also reveal more nuanced information on cellular

morphology. While Random Forest and SVM are robust classifiers, exploring other models, such as ensemble methods or deep learning approaches like fully supervised CNN classifiers, may also improve performance for handling subtle morphological variations. Variational autoencoders (VAEs) or transfer learning approaches might also reveal additional patterns in the data. The imbalance between DMSO and non-DMSO samples was a limiting factor in model performance. Future studies could consider sampling strategies, such as synthetic data generation or oversampling, to address this imbalance and improve model sensitivity to the minority class.