Building classifier for biomedical data

This is a computer vision project, which aims to build a classifier to distinguish between control and drug treated samples. Data comes from automated image processing pipeline, which corrects, normalizes and segments collected images to automatically detect and describe structures of interest with 200-1000+ attributes. To date, the dataset holds quantification data from 30 000+ images of both control and experimentally treated samples. The project will involve loading both images and corresponding quantification data into one database and testing various supervised learning algorithms for accurate binary classification. Unsupervised approaches will also be used to gain additional insights as to which features might contribute more strongly to the formation of two clusters and/or to find additional structure in the data. If successful, the project can be built upon by using existing image datasets from molecular treatments affecting other molecular pathways and incorporating established hierarchical ontologies to both name the classes and maintain known links between them. All in all, the project aims to lay a groundwork for applying data science methods for biomedical data to accelerate drug discovery.