ADS 599 – Proposal

The purpose of our Applied Data Science Capstone project is to automate flow cytometry data analysis using artificial intelligence to increase analytical throughput, increase objective classification accuracy, and identify insights that a human analyst might otherwise not observe.

Human analysts must observe and analyze various two-dimensional dot plots that depict characteristic markers such as fluorescence intensity, apply an arbitrary or manual gate to isolate specific types of cells or markers, and then potentially classify these different types of cells into their respective categories. Artificial intelligence can significantly reduce the manual classification process by applying unsupervised or clustering algorithms to be able to classify the dot plots with the cell type category that they best align with. This automation should result in increased throughput, more efficient clinical trials and experimentation, as well as potentially bring forth insights on multidimensional data that may be otherwise visually difficult for a human analyst to detect and gate in two-dimensional space.

Flow cytometry is a critical biochemical instrument that is used in the advancement of medicinal drug discovery, medicinal chemistry, and the overall pharmaceutical sciences. However, data analysis remains as a labor-intensive process that can be made significantly more efficient with artificial intelligence-based automation and machine learning algorithms. Exploring a new way to increase analytical throughput toward finding new medicines is a worthwhile and meaningful endeavor for data scientists to potentially discover life-saving medicines.

The team plans to gather the 1.5 gigabyte Flow Cytometry Standard files locally and only push resulting notebook, analytic, or automation modules of the project to GitHub in order to preserve storage capacity.

Three data science objectives are related to each business objective. First, we aim to identify unstructured clusters of cells using machine learning algorithms such as K-means Clustering, Density-Based Spatial Clustering of Applications with Noise, or Hierarchical Clustering in less time than it takes for a human to perform gating operations. Second, we aim for this process to classify the clusters within dot plots across different types of phenotypes with respective accuracy, recall, and specificity statistics. Finally, the team aims to classify at or above the industry threshold required for manual analysts to perform flow cytometry.

Our planned methodology involves gathering open-source research data from FlowRepository (n.d.). Then, we intend to preprocess the FCS files using the FlowCal package to parse and perform data transformation with standard packages including NumPy and MatPlotLib. Next, we will begin phenotype identification to determine the appropriate number of phenotypes and how to automate this process. Lastly, once the number of phenotypes is determined for a given dot plot, we will apply the optimal machine learning algorithm to classify the dot plots into phenotypes and then evaluate the results by accuracy.

Currently, real-world impact is to be able to affordably and efficiently perform automated biochemical analysis on a manual process, which would result in a significant reduction in costs and labor inputs to perform flow cytometry. This application may be implemented in either a Flask application or a Power BI or Tableau dashboard.

Finally, the team’s overall intent is to find a cost-effective or open-source alternative to automate analysis otherwise performed by existing industry programs such as Cytobank and FlowJo. Currently, both of these existing programs are the industry incumbents in terms of performing software-based analysis, however they do require paid licenses and subscriptions in order to be used.