ADS 599 – Proposal

Our Applied Data Science Capstone project aims to automate flow cytometry data analysis using artificial intelligence. This initiative seeks to enhance analytical throughput, deliver consistent and highly accurate classification across diverse single-cell samples, reduce the cost of AI-assisted gating, and uncover insights that might be overlooked by human analysts.

Flow cytometry is an important biochemical technique in drug research and clinical trials, empowering scientists to characterize diverse cell populations and elucidate mechanisms of action. Traditionally, manual flow cytometry analysis involves establishing density plots of cellular markers. Researchers then apply gates to isolate specific subpopulation and quantify the resulting groups. However, as the complexity and volume of analyses grow with the increasing number of fluorochromes used, the limitations of manual analysis become apparent. This escalating complexity underscores the pressing need for automated assistance in flow cytometry data interpretation, enabling more efficient, accurate, and comprehensive insights into cellular behavior and drug responses.

Personal experience in analyzing hundreds of Flow Cytometry Standard (FCS) files has revealed that the process is labor-intensive, time-consuming, and susceptible to subjective interpretation. This inefficiency creates a significant bottleneck in data analysis which then highlights the urgent need to streamline and optimize the workflow. We hypothesize that the application of artificial intelligence-based automation and machine learning algorithms will significantly reduce analysis time and improve consistency compared to manual methods. This approach offers a promising solution to the challenges of achieving reliable high-throughput results, with the goal of testing non-conventional cellular patterns. By leveraging AI and machine learning, we can expedite the delivery of results, paving the way for groundbreaking treatments and therapies that could significantly enhance patient care and outcomes.

Our data science objectives are closely aligned with our overarching business goals which is focusing on enhancing efficiency and accuracy in flow cytometry analysis. Firstly, we aim to leverage advanced machine learning algorithms – such as K-means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), or Hierarchical Clustering – to identify unstructured cell clusters significantly faster than traditional manual gating operations. Secondly, we will rigorously evaluate the performance of our classification models, targeting an accuracy of ≥ 90% and utilizing cumulative ROC-AUC plots for comprehensive assessment. Lastly, we will refine and optimize our best-performing model by subjecting it to challenging scenarios. This includes compromised or non-conventional sample(s) that can enhance the model’s robustness and real-world applicability.

Our planned methodology involves extracting flow files from the open-source FlowRepository website. We will strategically reduce the fluorochrome panel from 23 to 8 key markers, focusing specifically on dendritic cell populations. This targeted approach allows more precise and efficient data processing. 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. The core of our approach involves developing robust machine learning models to identify and classify distinct dendritic cell phenotypes using a carefully curated training set. We will implement rigorous cross-validation techniques before final testing to ensure model reliability and generalizability. The last step will be performance evaluation.

By leveraging publicly available AI technologies, we will develop a robust flow classification model that performs exceptionally across diverse single-cell samples in high multiplex fluorescence cocktails. This innovation will significantly reduce costs and labor inputs while accelerating result delivery and minimizing human error through standardized methods. Our solution aims to impact the real world by facilitating faster drug development and clinical trials, advancing precision medicine with consistent phenotyping capabilities and valuable insights from high-dimensional data.

We plan to implement a user-friendly interface, either as a Flask platform or a dashboard in Power BI or Tableau. Ultimately, we aim to provide an open-source alternative to new proprietary software like Cytobank, which requires costly licenses. By offering a powerful and accessible tool, we seek to enhance laboratory workflows, democratize advanced flow cytometry analysis, and drive innovation in biomedical research.