**CytoAutoCluster: Semi-Supervised Deep Approach for Cytometry Data Analysis**

**Project Overview**

CytoAutoCluster is a machine learning initiative aimed at enhancing the identification of cell populations through deep clustering techniques on cytometry data. By implementing semi-supervised learning models, the project seeks to improve both the accuracy and interpretability of cell classification.

**Data Cleaning and Preprocessing**

* **Exclusion of Irrelevant Features**: Certain columns, such as *Event*, *Time*, *Cell\_length*, and identifiers, were removed from the dataset as they were deemed non-informative for analysis.
* **Handling Missing Data**: We addressed missing values by implementing mean imputation, replacing missing entries with the mean of each respective column.
* **Standardization**: Data standardization was applied using StandardScaler to normalize features, ensuring that each had a mean of 0 and a standard deviation of 1. This step prepares the data for PCA and t-SNE by giving each feature an equal weight.

**Exploratory Data Analysis (EDA)**

* **Skewness and Kurtosis Analysis**: Each feature's distribution was examined for skewness and kurtosis to understand data distribution. Skewness indicates the asymmetry, while kurtosis measures the "tailedness" of the data, which helped in identifying the need for transformations.
* **Correlation Analysis**: A correlation matrix was generated, helping to identify multicollinearity among features. This analysis provided insights into feature interdependencies, which was especially useful before applying PCA.

**Dimensionality Reduction Techniques**

* **Principal Component Analysis (PCA)**: PCA was performed on the standardized dataset to reduce dimensionality while retaining as much variance as possible. We visualized PCA results in 2D and 3D scatter plots.
* **t-Distributed Stochastic Neighbor Embedding (t-SNE)**: t-SNE was applied as a nonlinear dimensionality reduction technique, particularly useful for visualizing high-dimensional data in two or three dimensions. The t-SNE plots helped reveal clustering and separation patterns in the data, indicating potential underlying structures.

**Binary Masking and Data Augmentation**

* **Binary Masking**: Implemented to simulate missing data by randomly masking 30% of feature values.
* **Shuffling Feature Columns**: Added column value shuffling to introduce variability and prepare data for semi-supervised clustering.