CytoAutoCluster : Revolutionizing Cytometry with Deep Learning

---

1. Data Import and Initial Exploration

- **Dataset**: Analysis is based on the **Levine_32dim** dataset.
- **Steps Taken**:
  - Imported the dataset and confirmed its structure.
  - Conducted preliminary checks to identify missing values, duplicates, and data types.

---

2. Data Preparation and Cleaning

- **Column Refinement**:
  - Removed unnecessary columns like file_number, Time, and event_number. Standardized column names for consistency.
- **Outlier Management**: Identified and addressed anomalies using statistical methods (e.g., Z-scores).
- **Correlation Mapping**:
  - Created a heatmap to visualize relationships between features.
- **Statistical Insights**:
  - Calculated skewness and kurtosis for feature distributions.
  - Generated histograms and box plots for detailed visualization.

---

3. Dimensionality Reduction Techniques

- **Principal Component Analysis (PCA)**:
  - Reduced the dimensionality of the dataset while retaining maximum variance.
  - Created both 2D and 3D visualizations to interpret the results.

- **t-SNE Visualization**:
  - Applied t-SNE for non-linear dimensionality reduction.
  - Enhanced cluster visualization for better feature separation.

---

4. Data Augmentation and Segmentation

- **Binary Masking**: Introduced labeled (corrupted) and unlabeled (original) subsets for semi-supervised learning.
- **Data Splitting**:
  - Divided the labeled subset into training and testing datasets to prepare for autoencoder training.

---

5. Semi-Supervised Autoencoder Preparation

- **Objective**:
  - Prepare the dataset to train an autoencoder capable of learning from both labeled and unlabeled data.
  - Focus on reconstructing corrupted data and extracting meaningful patterns.

---

6. Future Post-Training Analysis

- **Encoded Data Usage**:
  - Perform further dimensionality reduction and visualization.
  - Utilize clustering techniques to classify cell populations.

Key Takeaways

The **CytoAutoCluster** workflow demonstrates the power of semi-supervised deep learning in cytometry data analysis. By integrating advanced dimensionality reduction techniques and a robust autoencoder framework, this project establishes a foundation for accurate and interpretable cell population identification.