CytoAutoCluster:Enhancing Cytometry withDeep Learning

1. Introduction and Background

1.1 Objectives

- Perform data preprocessing and exploratory analysis.
- Develop and test semi-supervised clustering algorithms for high-dimensional mass cytometry data.
- Evaluate clustering results against manually gated clusters.

1.2 Overview of CyTOF Data

- Mass cytometry (CyTOF) is a high-throughput single-cell analysis technique.
- Enables measurement of over 30 markers per cell, commonly used in immune profiling and biomarker discovery.
- Analytical challenges: High dimensionality and noise in CyTOF data.
- This project explores semi-supervised clustering with autoencoder-based feature learning to address these challenges.

---

2. Background and Motivation

2.1 Cytometry Data Analysis

- Cytometry measures cellular properties like size, complexity, and protein expression.
- Applications: Immunology, cancer research, and disease diagnostics.

2.2 Challenges

1. High dimensionality of data.
2. Noise introduced by biological and technical factors.
3. Limited availability of labeled data.

2.3 Motivation for CytoAutoCluster

- Combines semi-supervised learning to improve cluster precision and interpretability.

---

3. Dataset Description

3.1 Properties

- Total cells: **265,627**
- Markers: **32**
- Labeled cells: **39% (104,184 cells)**
- Unlabeled cells: **61% (161,443 cells)**
- Number of clusters: **14**

3.2 Source

- Dataset: *Levine_32dim.csv*

- Reference: Levine et al. (2015), publicly available on Cytobank.

3.3 Marker Details

- **Markers for Manual Gating:** CD3, CD4, CD7, CD8, HLA-DR, CD123, CD235a/b, etc.
- **Additional Markers:** CD10, CD45RA, CD56, etc.

---

4. Methodology

4.1 Data Preprocessing

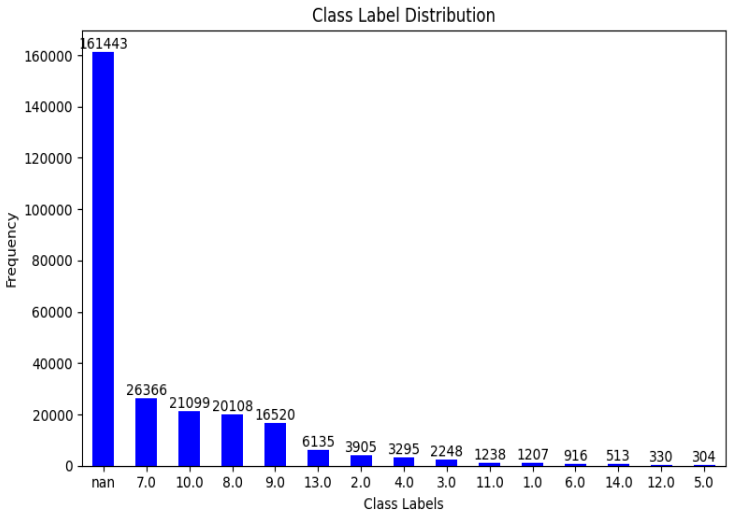
- **Normalization:** StandardScaler used to normalize feature distributions.
- **Exploratory Analysis:** Histograms and density plots analyzed marker distributions.
- **Data Masking:** Simulated partially labeled data for real-world scenarios.

4.2 Data Splitting

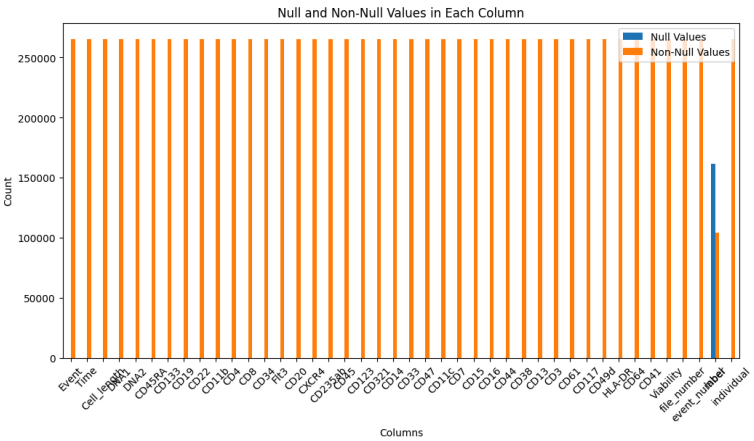
- Labeled data split into training and testing sets (70-30 split).

4.3 Data Exploration

- **Cluster Distribution Analysis:** Visualize cluster imbalances

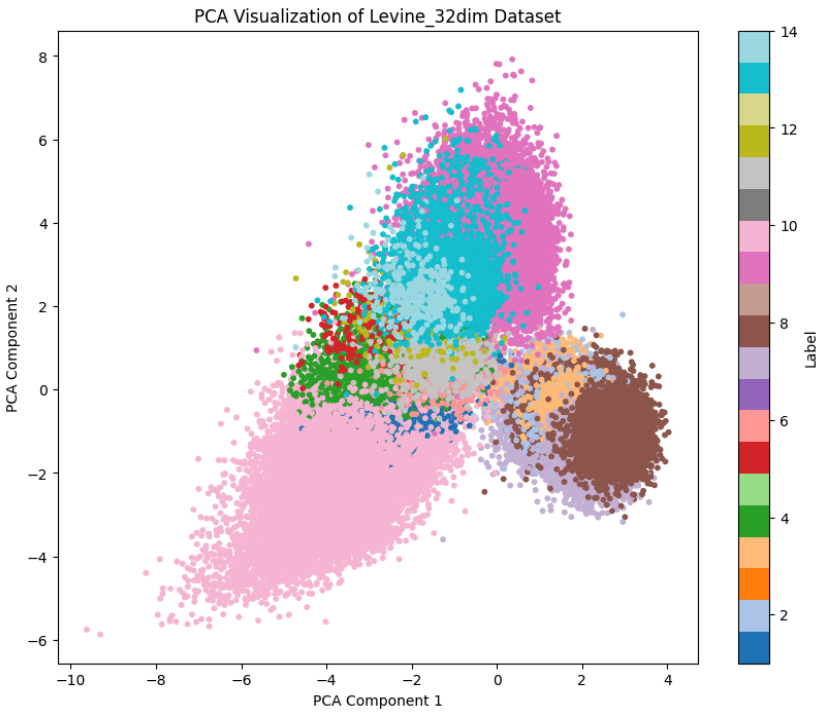


- **Null Value Analysis:** Examine null vs. non-null values in labels

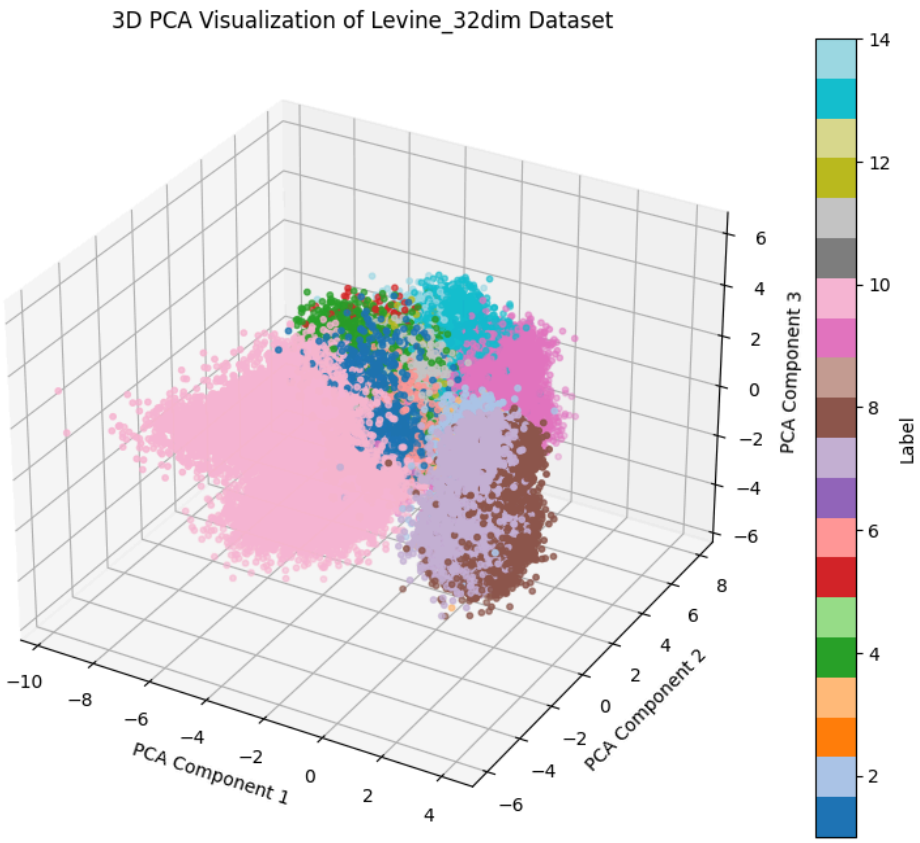


4.4 Clustering Techniques

- **Autoencoder-based Dimensionality Reduction:** Extract latent representations.
- **PCA:** It reduces dimensionality by retaining essential features.

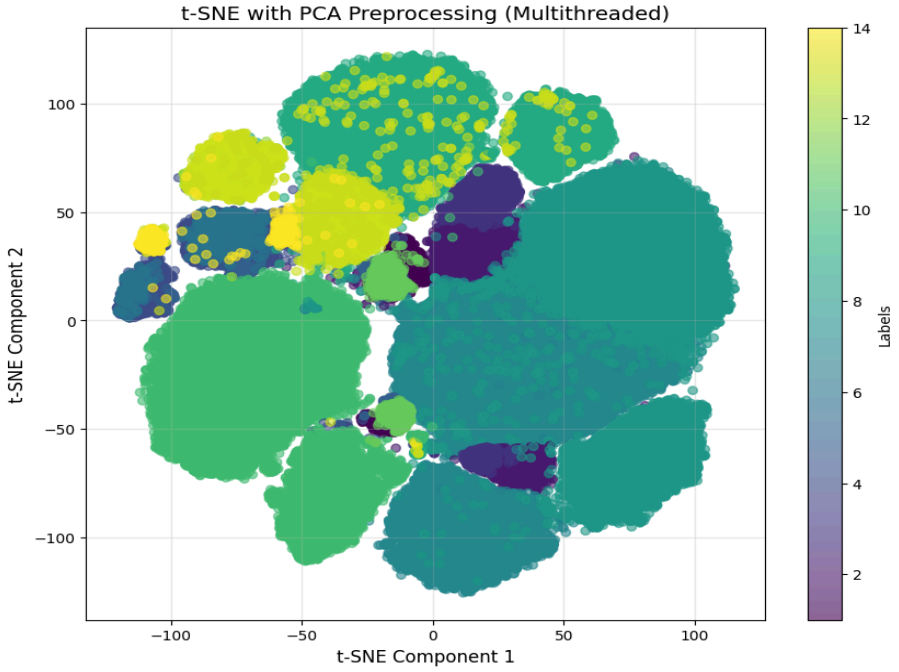


PCA 2D plot



PCA 3D Plot

- **t-SNE Visualization:** Identify clusters in lower-dimensional space



---

5. Implementation

5.1 Feature Engineering

- Label encoding for categorical labels.
- StandardScaler for feature scaling.

5.2 Semi-Supervised Learning

- Binary mask with probability **p_m = 0.5** introduced controlled corruption.
- Loss functions used:
  - **Binary cross-entropy loss** for mask prediction.
  - **MSE loss** for feature reconstruction.

5.3 Supervised Fine-Tuning and Testing

5.3.1 Logistic Regression

- **Purpose:** Interpretable model with probabilistic outputs.
- **Results:** Achieved a log loss of 0.0299 on test data.

5.3.2 XGBoost

- **Purpose:** Non-linear, ensemble-based model.
- **Results:** Achieved a log loss of 0.0039, highlighting improved performance.

5.4 Model Comparison

Model	Log Loss	Advantages
Logistic Regression	0.0331	Efficient, interpretable for high-dimensional data.
XGBoost	0.0040	Robust for non-linear relationships

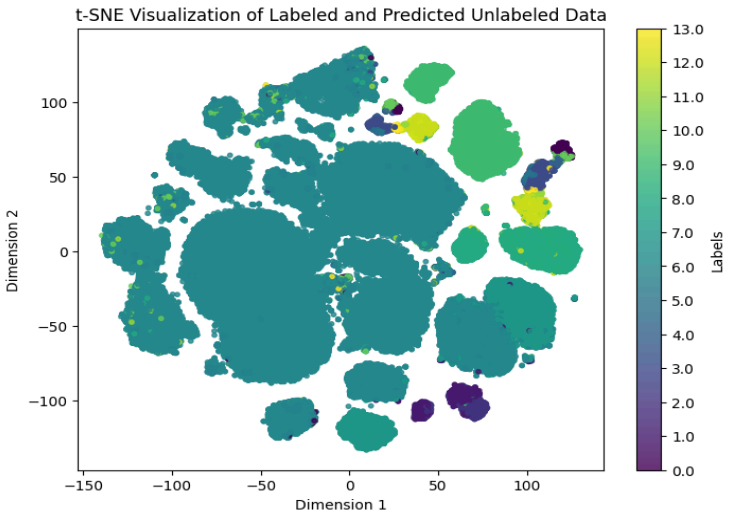
6. Results and Evaluation

6.1 Quantitative Metrics

- **Log Loss:** Evaluated clustering accuracy.
- **Accuracy:** Measured overall performance.

6.2 Visual Insights

- t-SNE plots highlight clear cluster separations



7. Gradio Interface for Visualization

7.1 Overview

- Provides a dynamic interface for visualizing t-SNE outputs and predicted labels.

Functions Implemented:

1. **Prediction Function:** Encodes unlabeled data and generates predictions.
2. **t-SNE Visualization:** Projects high-dimensional data into interpretable 2D space.
3. **Gradio Integration:** Processes subsets of data for interactive exploration.

8. Challenges and Solutions

- **Noisy Data:** Addressed with scaling, normalization, and random masking.
- **Label Imbalance:** Used semi-supervised learning and adjusted class weights.
- **High Dimensionality:** Employed dimensionality reduction techniques like autoencoders.

---

9. Future Work

1. Extend to multi-omics data (e.g., genomics, proteomics).
2. Apply to diverse cytometry datasets, including flow cytometry.
3. Enhance scalability with distributed computing.
4. Incorporate domain adaptation for cross-laboratory variability.

---

10. Conclusion

The CytoAutoCluster framework:

- Effectively clusters high-dimensional cytometry data with limited labels.
- Improves interpretability for rare and complex cell subtypes.
- Advances diagnostics by enabling better classification of cellular phenotypes.

---

11. References and Acknowledgments

1. Levine, J. H., et al. (2015). Data-Driven Phenotypic Dissection of AML. Cell, 162, pp. 184–197
2. Publicly available cytometry datasets (e.g., Kaggle).

Demo Link: Cytoautocluster.mp4
