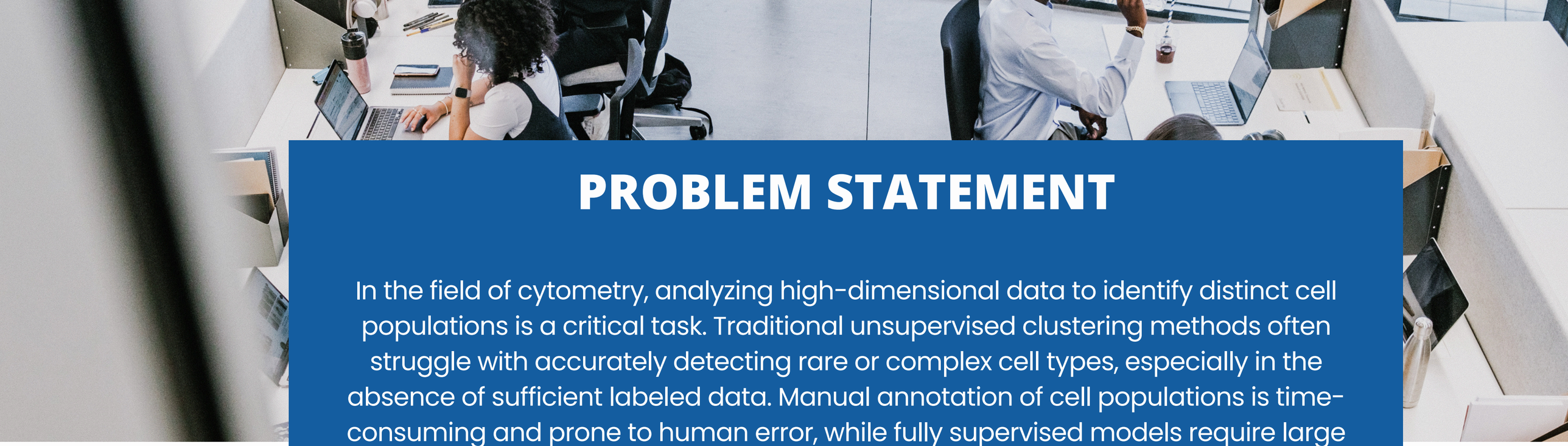
CytoAutoCluster
A Semi-Supervised Deep Approach For Cytometry Analysis

By: HARI PRIYA NAGANDLA

Overview

- PROBLEM STATEMENT 01
- CHALLENGES 02
- INTRODUCTION 03
- OBJECTIVE 04
- DATA EXPLOITATION 05
- METHODOLOGY 06
- MODEL DESCRIPTION 07
- RESULT AND EVALUATION 09
- GRADIO INTERFACE 10



PROBLEM STATEMENT

In the field of cytometry, analyzing high-dimensional data to identify distinct cell populations is a critical task. Traditional unsupervised clustering methods often struggle with accurately detecting rare or complex cell types, especially in the absence of sufficient labeled data. Manual annotation of cell populations is time-consuming and prone to human error, while fully supervised models require large amounts of labeled data, which may not always be available.

CytoAutoCluster aims to address these challenges by leveraging semi-supervised deep clustering techniques to improve the identification and interpretation of cell populations from cytometry data.

.

CHALLENGES

01

HIGH DIMENSIONALITY

THOUSANDS OF FEATURES

SOLUTION: FEATURE SELECTION

02

NOISE IN DATA

Due to instrument variability or biological factors

SOLUTION : DATA AUGMENTATION

03

LIMITED LABELED DATA AVAILABILITY

RARE CELL POPULATIONS

SOLUTION: SEMI-SUPERVISED LEARNING

Introduction

- This work demonstrates a semi-supervised learning approach using an encoder model to process unlabeled data and predict class labels.
- The model leverages dimensionality reduction techniques like t-SNE to visualize high-dimensional data.
- Gradio Interface is used to allow users to interact with the model and explore predictions.
- Utilize an encoder to generate features from unlabeled data, then apply a trained supervised model (predictor) to classify those features.
- The results are visualized using t-SNE for clustering of labeled and predicted unlabeled data.

Objective

- Utilize an encoder to generate features from unlabeled data, then apply a trained supervised model (predictor) to classify those features.
- The results are visualized using t-SNE for clustering of labeled and predicted unlabeled data.

.

Key-Steps

Step 1	Step 2	Step 3	Step 4
Preprocessing of unlabeled data	Prediction using a pre-trained encoder and classifier.	Visualization of predictions via t-SNE.	Interactive Exploration using Gradio.

DATA EXPLOITATION

Data Sources:

- Labeled Data: Used for training the predictor model.
- Unlabeled Data: Used to generate predictions through the encoder and predictor.

01 SELECT RIGHT DATASET (LIKE NUMBER OF FEATURES, FEATURE TYPES, CHECKING FOR BIASES)

02 CONDUCT EDA TO CONFIRM DATASET SUITABILITY

03 DATASET : Levine32Dimensional (Kaggle)
Labeled : 39%
Unlabeled : 61%

04 Challenges : High dimensionality, large proportion of missing labels, 14 clusters in label

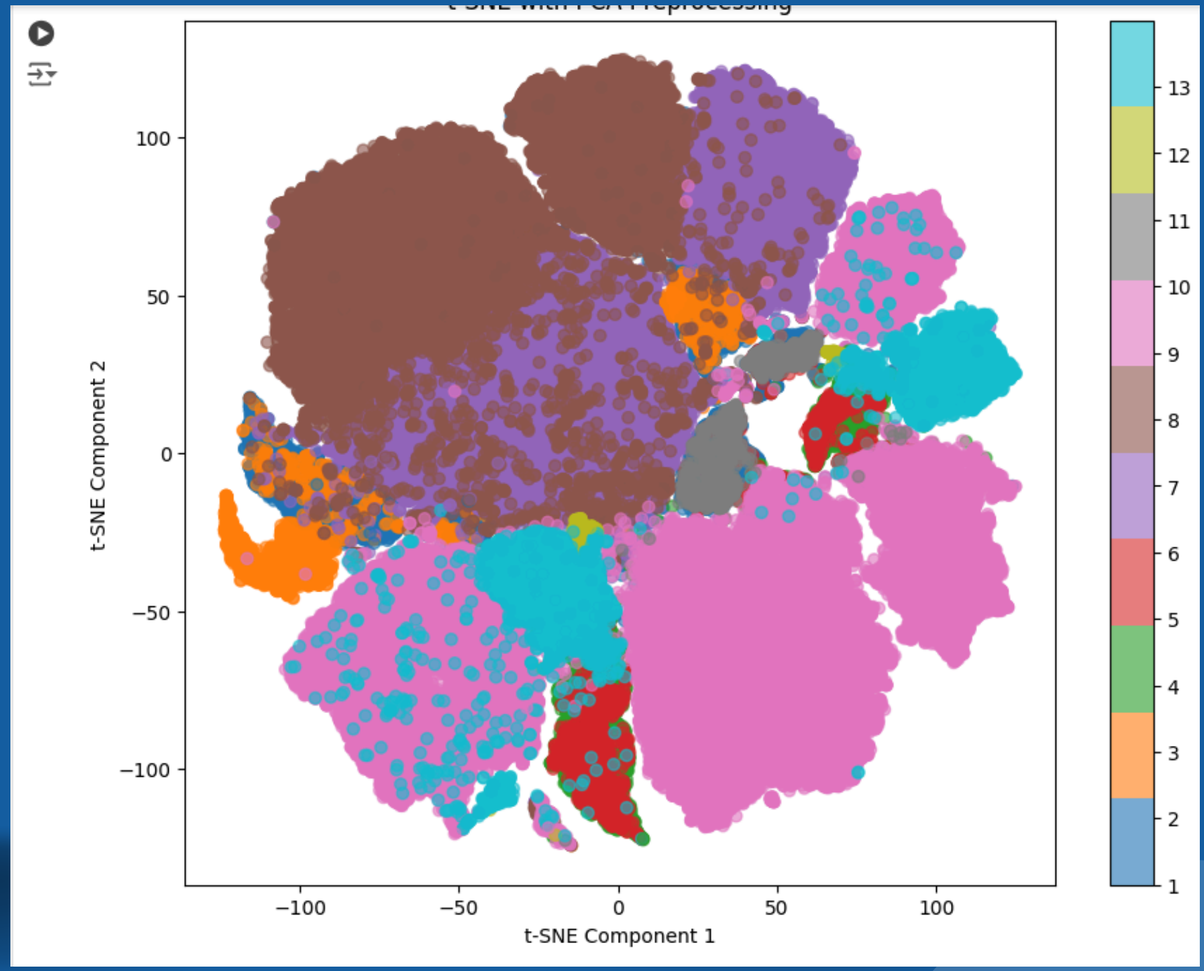
METHODOLOGY

DATA PREPROCESSSING

- NORMALIZATION: Standardized feature distributions using StandardScaler
- Exploratory Analysis: Visualized marker distributions through histograms and density plots
- DataMasking: Simulated partially labeled data for real-world clustering scenarios

Clustering Technique:

- Autoencoder-based Dimensionality Reduction: Extracted compact latent representations
- t-SNEVisualization: Identified clusters in two or three dimensions.



MODEL DESCRIPTION

SEMI-SUPERVISED LEARNING

- self-supervised encoder for feature extraction
- supervised fine-tuning with logistic regression and XGBoost

TRAINING DETAILS

- Batch size, epochs, optimizers
- handling Label Imbalances with class weights

MODEL DESCRIPTION

-> model.summary()

Model: "functional"

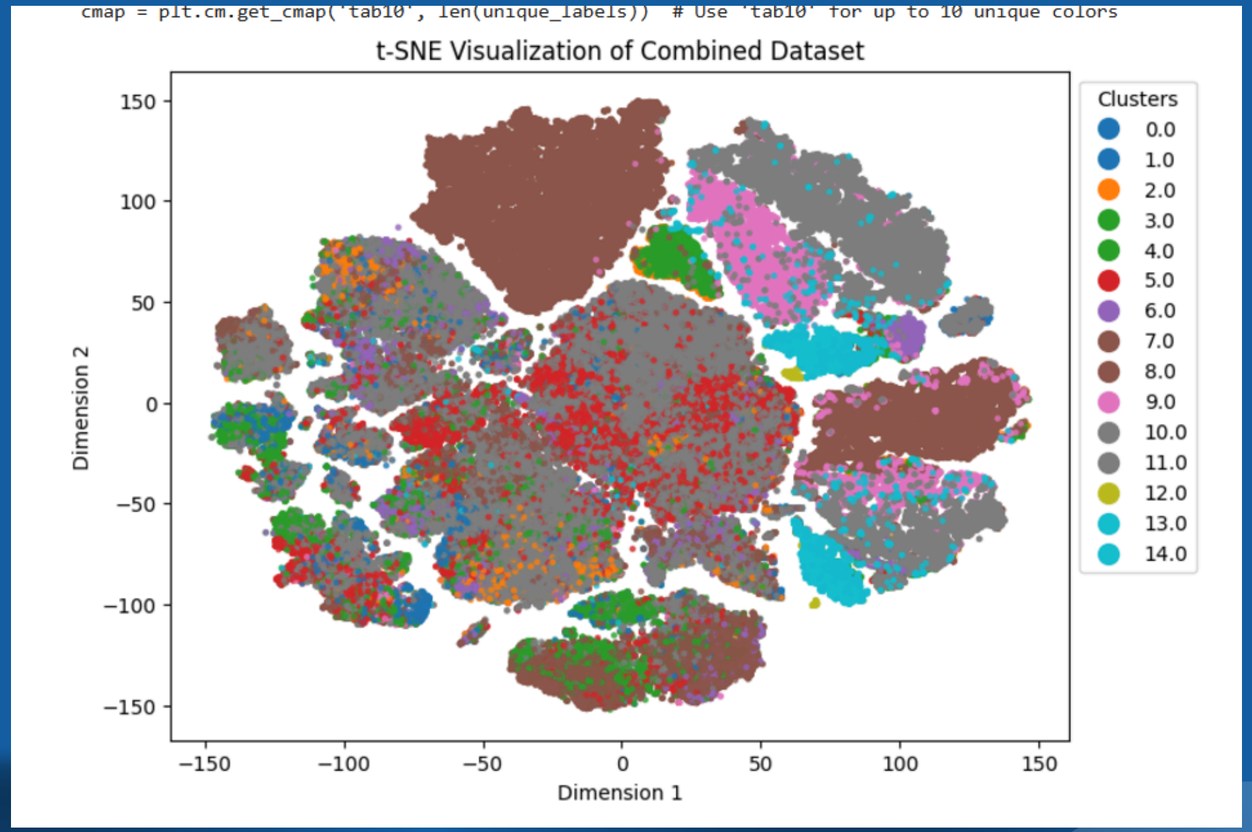
Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 36)	0	-
dense (Dense)	(None, 36)	1,332	input_layer[0][0]
mask_estimation (Dense)	(None, 36)	1,332	dense[0][0]
feature_estimation (Dense)	(None, 36)	1,332	dense[0][0]

Total params: 7,994 (31.23 KB)
Trainable params: 3,996 (15.61 KB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 3,998 (15.62 KB)

RESULT AND EVALUTION

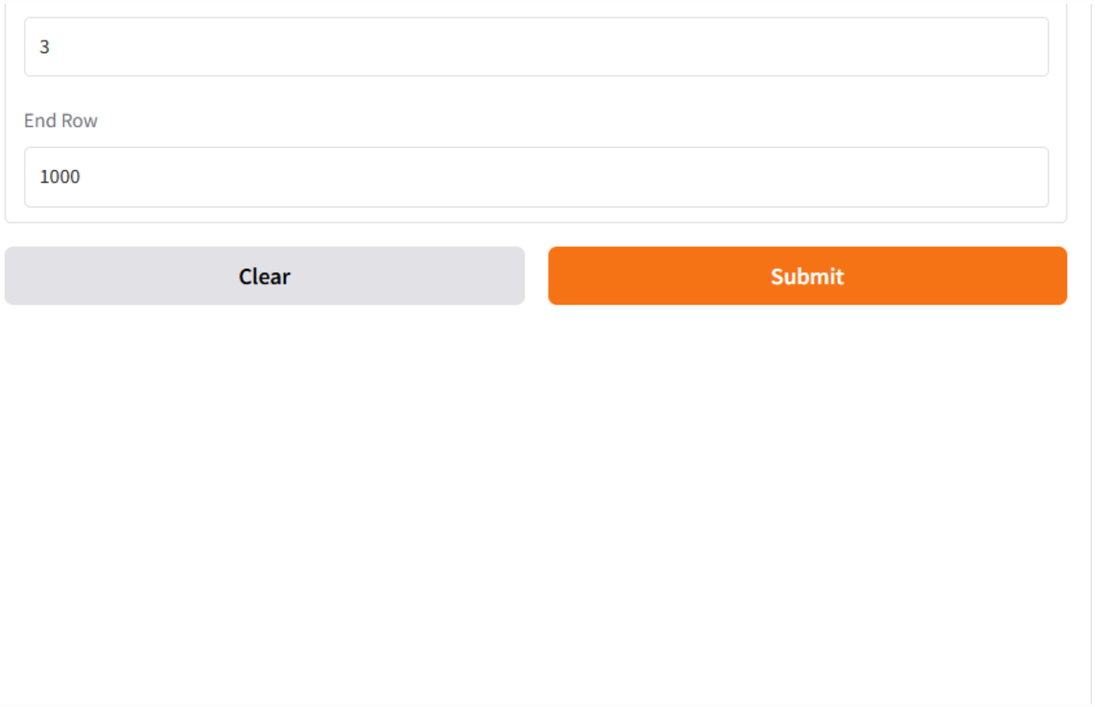
This work demonstrates

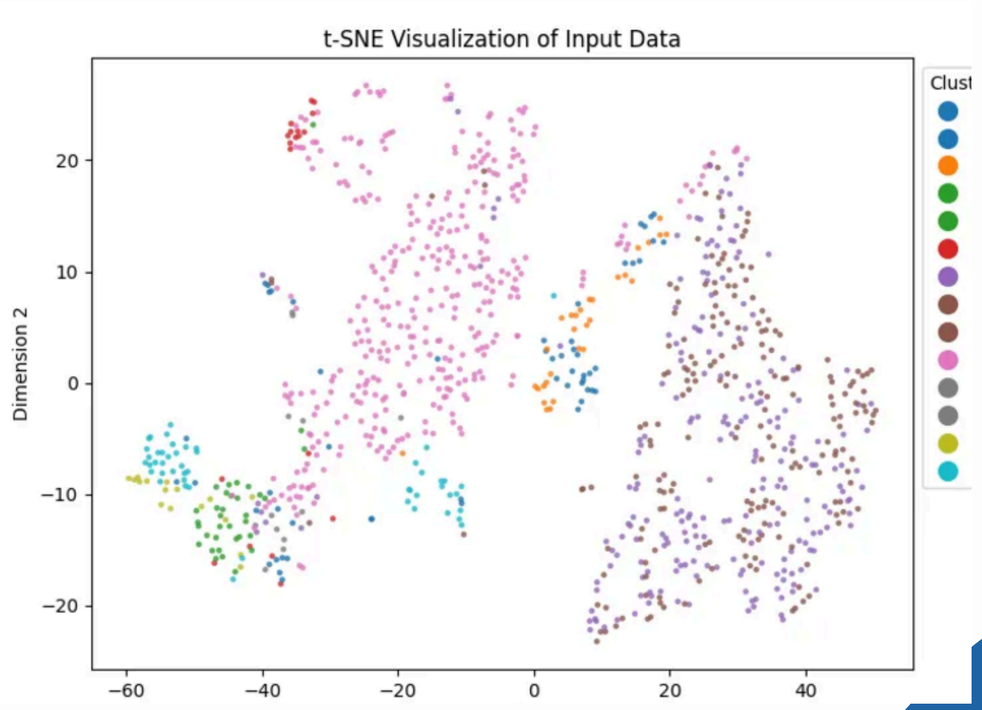
- Quantitative Metrics:
  - Logistic regression Log Loss: 0.0649
  - XGBoost Log Loss : 0.0073
  - Self-supervised Model Accuracy 93.5%
  - AUROC : 99.09%
  - VISUAL INSIGHTS : t-SNE Plot showcasing cluster seperation



GRADIO INTERFACE

Accepts start and end rows for the subset of unlabeled data. Processes data, predicts labels, and returns both a t-SNE plot and predictions in a tabular format





ACCESS TO THE LIVE DEMO

CLICK HERE

THANK YOU!