Final Project: Cell Type Annotation Using Machine Learning

Ethiopia Woldemarriam V00977761

Joseph Foster V00985606

Hazza Sawant V01017452

Ian Sam V00990790

Stat 469 A01, Professor Xuekui Zhang

April 13th, 2025

Introduction

In this project, we address the challenge of cell type annotation by approaching it as a multi-class classification problem using machine learning techniques. The task involves working with two datasets: a labeled training dataset used to develop and fine-tune models and an unlabeled test dataset, for which the goal is to accurately predict cell type labels based on gene expression data.

The objective is to build a model that performs well on the training data while maintaining strong generalization capabilities when applied to unseen samples. To establish a reference point, we begin with a Random Forest (RF) model as our **baseline model**. The Random Forest model provides a critical benchmark against which we will evaluate more advanced models to better fit our data.

Baseline Model: Random Forest

Model Setup

We implemented a Random Forest classifier with the following configuration:

- 100 decision trees (`n_estimators=100`)
- Maximum depth of 10 (`max_depth=10`) to prevent overfitting
- Stratified train-validation split (80% training, 20% validation) to ensure balanced class representation

Cell Type	Precision	Recall	F1-score	Support
1	1.00	0.78	0.88	90
2	0.95	1.00	0.97	4954
3	1.00	0.16	0.28	166
4	1.00	0.08	0.14	78
5	1.00	0.08	0.14	13
6	0.00	0.00	0.00	4
7	0.86	0.97	0.91	740

8	1.00	0.14	0.25	112
9	0.00	0.00	0.00	13
10	1.00	0.45	0.62	33

Table 1: Random Forest results

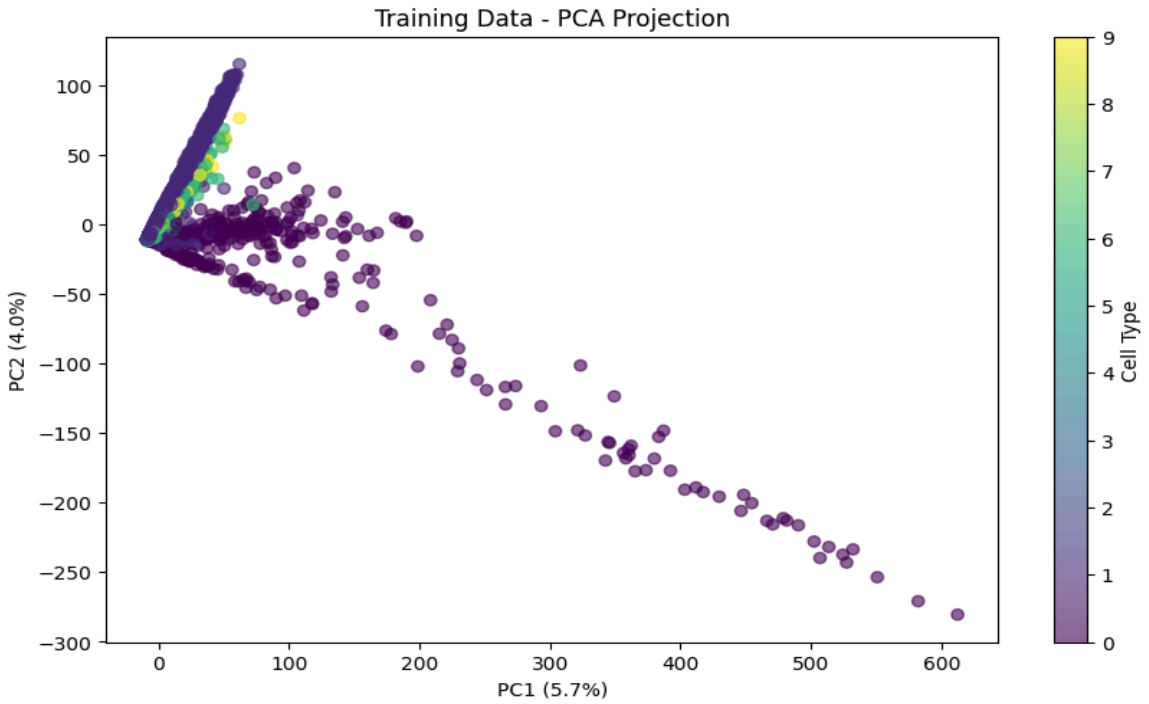

Figure 1: PCA projection for the Baseline Mode

Baseline model achieved:

- Overall weighted accuracy = 94%
- Weighted F1-score = 0.91
- Macro F1-score = 0.42

Our Random Forest baseline achieved strong overall performance (94% accuracy, weighted F1=0.91) but revealed critical limitations in handling class imbalance. While the model excelled at predicting majority cell types, it performed poorly on rare classes, with several minority types showing F1-scores below 0.3. This disparity highlights two key challenges: the model's bias toward dominant populations due to extreme class imbalance and insufficient signal capture for low-abundance cell types with default parameters. To address these limitations, our

future models must aim for a minimum recall of 50% across all classes, a macro F1-score above 0.65 and sustain a weighted F1-score greater than 0.90 to ensure both fairness and overall reliability. In *figure 1,* we can observe a lot of overlap between cell types and represent that the baseline method may have difficulty distinguishing meaningful clusters.

Methodology and Model Development

Exploratory Model: K-means Clustering

We initially explored K-means clustering under the assumption that gene expression profiles may form distinct natural groupings corresponding to cell types. The following configuration was used:

Model Setup

- PCA reduction to 50 components to reduce dimensionality and noise
- K-means with 10 clusters (based on known cell types)
- Cluster-to-label mapping based on the majority label within each cluster

Cell Type	Precision	Recall	F1-score	Support
1	0.99	0.55	0.71	359
2	0.82	1.00	0.90	19812
3	0.00	0.00	0.00	664
4	0.00	0.00	0.00	311
5	0.00	0.00	0.00	52
6	0.00	0.00	0.00	17
7	0.99	0.15	0.26	2960
8	0.00	0.00	0.00	447
9	0.00	0.00	0.00	53
10	1.00	0.51	0.68	133

Table 2: K-means clustering results

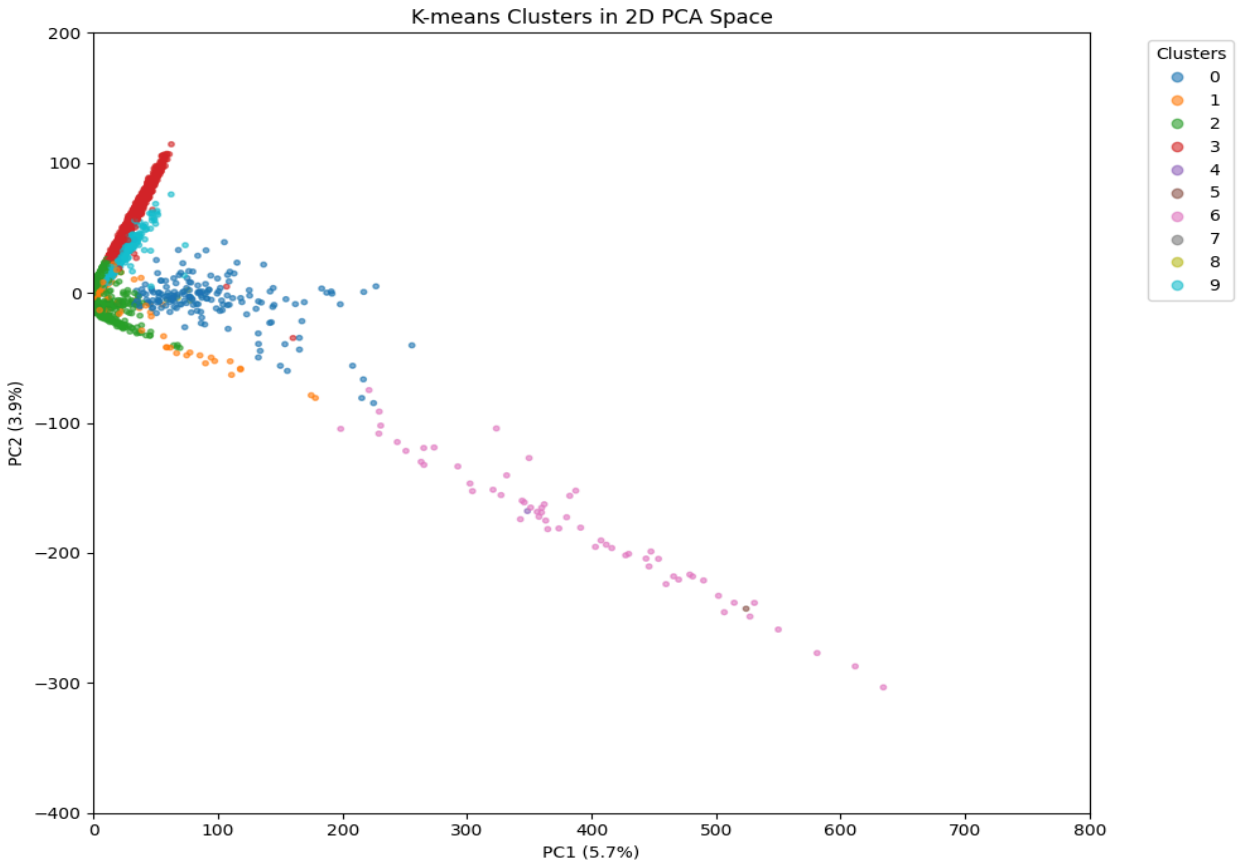

Figure 2: PCA Projection for K-means clusters

The K-means Model Achieved:

- Accuracy = **82.7%**
- Weighted F1-score = **0.77**
- Macro F1-score = **0.25**

Although K-means achieved a modest overall accuracy of 83% and a weighted F1-score of 0.77, a deeper look at the class-wise results revealed severe limitations. While the dominant class (2.0) was predicted well (F1 = 0.90), the model completely failed to identify numerous rare cell types such as 3.0, 4.0, 5.0, 6.0, 8.0, and 9.0, all of which had F1-scores of 0.00. The macro F1-score dropped to just 0.25, indicating a strong bias toward majority classes. These results highlighted the inability of unsupervised methods to handle class imbalance or capture subtle expression patterns. To address these challenges, we shifted to a supervised gradient boosting approach, **XGBoost**, which could incorporate label information, balance class weights, and optimize performance across all cell types more effectively. In *figure 2,* we observe clearer

separation between several clusters. However, significant overlap remains in high-density regions, and rare cell types are not distinctly separated.

Supervised Model: XGBoost

To overcome the limitations of K-means and improve classification across all cell types, we implemented XGBoost, a supervised gradient boosting model known for its strong performance on high-dimensional, structured datasets and its ability to effectively handle imbalanced classes through weighting and regularization.

Model Setup

- Set n_estimators = 500
- Used learning_rate = 0.05
- Enabled early stopping rounds = 20

Cell Type	Precision	Recall	F1-score	Support
1	1.00	0.92	0.96	90
2	0.98	0.99	0.99	4954
3	0.90	0.69	0.78	166
4	0.85	0.56	0.68	78
5	1.00	0.69	0.82	13
6	1.00	1.00	1.00	4
7	0.97	0.97	0.97	740
8	0.90	0.89	0.90	112
9	0.92	0.92	0.92	13
10	0.94	0.88	0.91	33

Table 3: XGBoost results

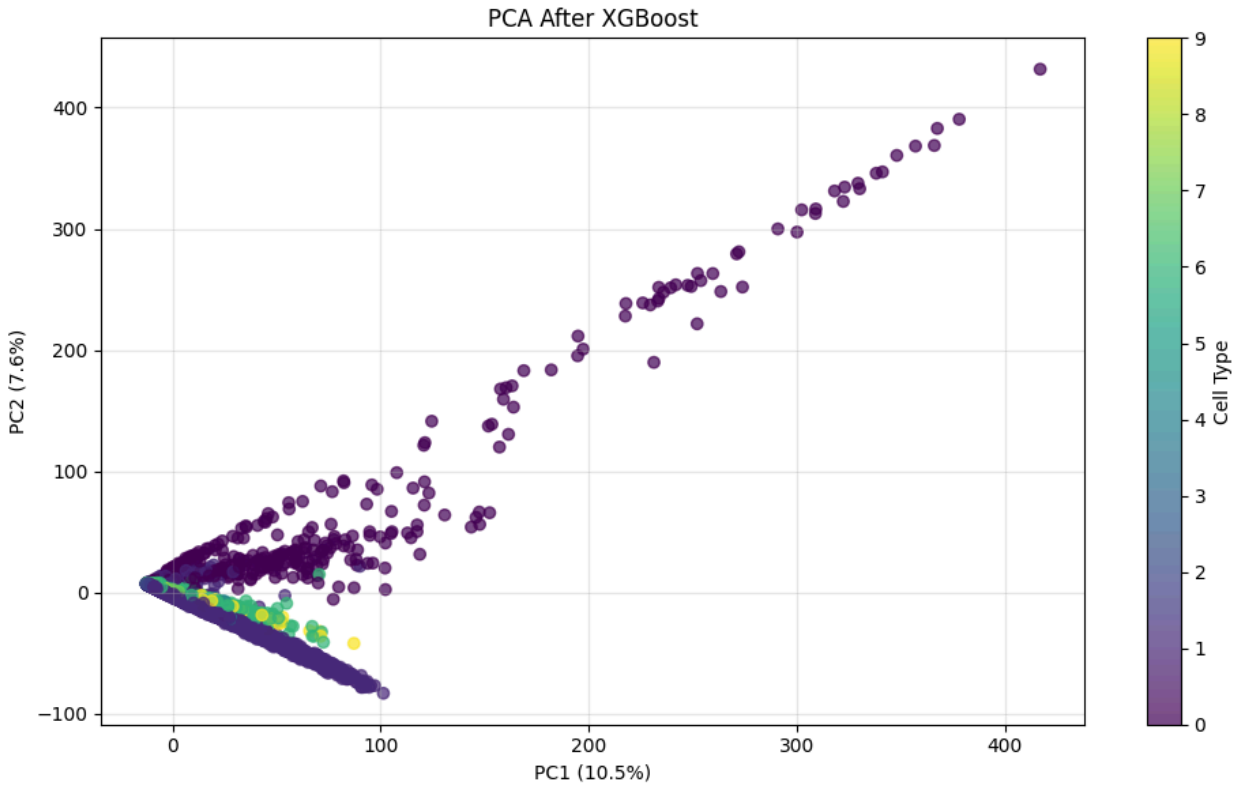

Figure 3: PCA projection for XGBoost

The XGBoost Model Achieved:

- Accuracy = **97.3 %**
- Macro F1 = **0.892**
- Weighted F1 = **0.972**

Although XGBoost delivered a strong overall performance with an accuracy of 97.3% and a weighted F1-score of 0.972, class imbalance continued to affect the model's ability to consistently classify rare cell types. While dominant classes such as 2.0 and 7.0 were predicted with near-perfect precision and recall, minority classes like 4.0 and 5.0 still lagged with F1-scores of 0.68 and 0.82, respectively. The macro F1-score, although much improved from K-means, remained at 0.892, which indicates uneven performance across classes. To further enhance recall and representation for underrepresented cell types, we introduced SMOTE to synthetically balance the training dataset. In *figure 3,* we see increased spread along the principal components, indicating that the model has learned more complex, non-linear patterns. However, despite this improved feature representation, rare cell types remain visually clustered with dominant ones, highlighting the model's ongoing struggle with class imbalance, especially for low-support classes.

Supervised Model: XGBoost + SMOTE

In our analysis, SMOTE was introduced to complement XGBoost to directly address the challenge of severe class imbalance present in the training data. Although XGBoost is a highly effective supervised learning algorithm, its predictive performance can be compromised when minority classes are significantly underrepresented. SMOTE mitigates this problem by synthesizing new samples for these minority classes through the interpolation of feature space among existing samples, which provides a more balanced dataset. This allows XGBoost to learn more representative patterns for all cell types, enhancing its ability to detect and correctly classify rare cell populations without sacrificing overall accuracy.

Model Setup

- Set n_estimators = 500
- Used learning rate = 0.05
- Enabled early stopping rounds = 20
- We applied **targeted SMOTE** to minority classes (4.0 and 5.0)

Cell Type	Precision	Recall	F1-score	Support
1.0	1.00	0.97	0.98	90
2.0	0.99	0.99	0.99	4954
3.0	0.86	0.86	0.86	166
4.0	0.78	0.65	0.71	78
5.0	1.00	0.85	0.92	13
6.0	0.80	1.00	0.89	4
7.0	0.97	0.98	0.97	740
8.0	0.87	0.91	0.89	112
9.0	0.92	0.92	0.92	13
10.0	0.94	0.88	0.91	33

Table 4: XGBoost and SMOTE results

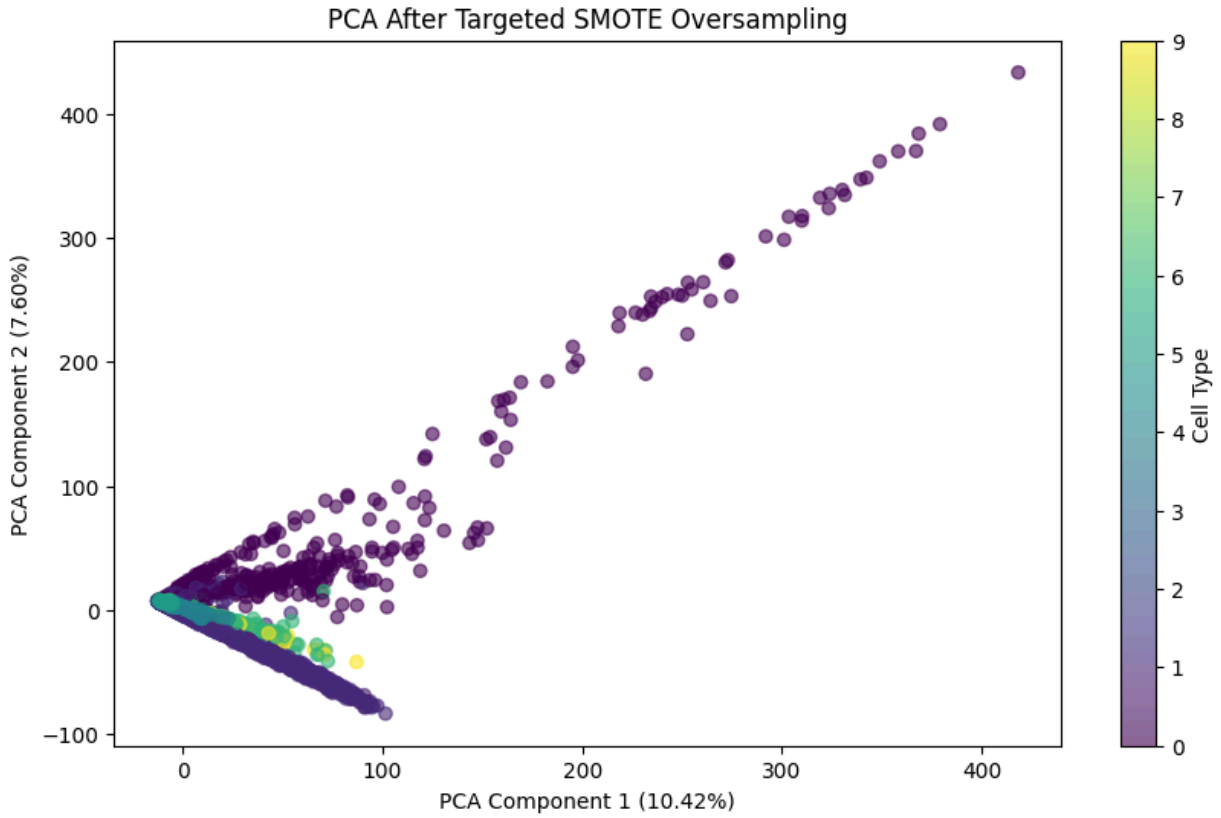

Figure 4: PCA projection for the Final model

The XGBoost + SMOTE Model Achieved:

- Accuracy = **97.7 %**
- Macro F1 = **0.904**
- Weighted F1 = **0.977**

The XGBoost + SMOTE model achieved the strongest performance across all evaluation metrics, with an accuracy of 97.7%, a macro F1-score of 0.904, and a weighted F1-score of 0.977. These results demonstrated substantial improvements in fairness across cell types. Compared to the original XGBoost model, applying SMOTE notably enhanced the recall and F1-scores of minority classes such as 3.0, 4.0, 5.0, and 6.0. For instance, cell type 3.0 improved from an F1-score of 0.78 to 0.86, and cell type 5.0 from 0.82 to 0.92. These gains contributed to a higher macro F1-score, indicating more balanced performance across classes. While XGBoost alone struggled to sufficiently learn from rare cell populations, the inclusion of synthetic minority samples allowed the model to capture these patterns more effectively. This combination ensured that both high-frequency and low-frequency cell types were accurately classified, making XGBoost + SMOTE the most reliable and equitable model in our analysis. In *Figure* 4,

After applying SMOTE, the PCA projection shows improved dispersion of minority cell types, indicating better representation and separability. Overall, applying SMOTE after XGBoost improved minority class representation and separability in PCA space, leading to more balanced learning and stronger classification performance across all cell types.

Novelty of Approach

Our approach in this report introduces Extreme Gradient Boosting (XGBoost). XGBoost uses labeled training data to find non-linear relationships between the cell types. The K-means method does not use labels and, therefore, cannot understand what cell types are and simply uses distances between samples to form clusters. XGBoost makes up for those limitations in the K-means algorithm and is ideal for the data used in this project. In addition to XGBoost, we also used the Synthetic Minority Oversampling Technique (SMOTE) in our project. This method generates synthetic data for underrepresented classes. This directly addresses an issue of class imbalance. In our cell type annotation, this greatly improves the performance on minority cell types. The improvement in our accuracy and F1 scores after the inclusion of these techniques not discussed in class–such as XGBoost and SMOTE–demonstrates the novelty of our approach.

Conclusion

This project explored the challenge of cell type annotation using various machine learning approaches, beginning with a Random Forest baseline and progressing through K-means clustering, XGBoost, and ultimately XGBoost enhanced with SMOTE. While the Random Forest model provided strong initial performance, it struggled with rare cell types due to class imbalance. K-means, as an unsupervised method, failed to capture meaningful patterns for minority classes. In contrast, XGBoost delivered significant improvements by leveraging labeled data and handling complex relationships. The addition of SMOTE further boosted performance, especially for underrepresented classes, leading to our highest accuracy and most balanced classification results. This final model, **XGBoost + SMOTE**, delivered the most equitable and accurate results, with a accuracy of 97.7% and a weighted F1-score of 0.977.
