Cell Type & Disease Status Classification of scRNA-seq Cancer Microenvironment Data

Yazan Moakkit
Electrical Engineering
ymoakkit@terpmail.umd.edu

Luke Stadler
Computer Engineering
lstadler@terpmail.umd.edu

Pranav Chandar Sridar
Computer Engineering
pranav25@terpmail.umd.edu

Brian Wu
Computer Engineering
bwu32@terpmail.umd.edu

***Abstract*—Effective cancer treatment requires understanding tumors at single-cell resolution. We present a hierarchical two-layer AdaBoost pipeline for classifying cell types and disease status in synthetic single-cell RNA sequencing (scRNA-seq) data from tumor microenvironments. Layer 1 achieves 100% accuracy in distinguishing Cancer, T_Cell, and Fibroblast populations using ensemble majority voting across five diverse base learners. Layer 2 predicts disease status (Tumor vs. Healthy_Control) for non-cancer cells, achieving 87.5% overall accuracy and 81.4% accuracy on non-cancer samples specifically. Our key finding demonstrates that weak learners (decision stumps with max_depth=1) significantly outperform complex models (Linear SVM) when combined with AdaBoost, achieving 87.4% accuracy versus 71.6%. The pipeline successfully addresses severe class imbalance (76% tumor-associated cells) through strategic train-test splitting and ensemble diversity, providing a principled framework for hierarchical classification in computational biology.**

***Index Terms*—single-cell RNA sequencing, AdaBoost, ensemble learning, cancer classification, tumor microenvironment, machine learning**

I. Introduction

Effective treatment for cancer depends on the understanding of tumors at single-cell resolution instead of as homogenous masses. Advances in the research of single-cell RNA sequencing (scRNA-seq) make it possible to profile thousands of individual cells within a tumor, capturing the diverse gene expression programs of malignant cells [1], [2]. However, creating reliable patterns from noisy expression profiles and mapping each cell to a biologically meaningful state can be a time-consuming, manual process, especially if there are thousands of samples to be analyzed. Computational pipelines that can accurately classify cellular subtypes within the tumor and determine whether they may be malignant or benign can be created through machine learning principles [3]. Synthetic datasets that mimic realistic tumor microenvironment (TME) structure provide insight into developing predictive patterns for how tumors evade immunity, remodel surrounding tissue, and respond to treatment [4].

This project aims to classify individual cells in a synthetic scRNA-seq dataset into three cell types (Cancer, T_Cell, Fibroblast) and to determine whether non-cancer cells originate from tumor or healthy control tissue. Each of the 3,000 simulated cells is described by expression levels of five biologically meaningful genes, an inflammation-related pathway score, and a low-dimensional UMAP coordinate, creating a multi-dimensional classification problem where relevant patterns must be extracted from complex gene activity profiles. There is an urgent clinical need for accurate and scalable tools to support cancer diagnosis and characterization, given the large annual cancer burden and limitations of manual, error-prone diagnostic workflows [5]. To tackle this, our project proposes a hierarchical, two-layer pipeline that first predicts cell type using a diverse set of base learners combined through majority voting, then applies AdaBoost classifiers to distinguish Tumor versus Healthy_Control within non-cancer cells, leveraging meta-features from the first layer to improve disease status prediction.

II. Dataset Description

The scRNA-seq: Cancer Microenvironment Classification dataset simulates a TME with three embedded cell types (Cancer, T_Cell, Fibroblast) and tumor tissue biology. It is composed of 3,000 cells and uses a log-normal model to mimic real data sparsity and variability. There are five key genes (Gene A–Gene E) and a pathway score derived from inflammatory activity, as well as a UMAP_1 coordinate for dimensionality reduction. Gene A is highly expressed in Cancer cells, Gene B marks immune activity in T_Cells, Gene C reflects stromal/fibrotic signaling in Fibroblasts, Gene D represents a therapeutic target with high expression in Cancer but low in T_Cells, and Gene E serves as a stable housekeeping reference. Ground truth ties Cancer cells directly to Tumor tissue, while non-cancer cells can be labeled as Tumor-associated or Healthy_Control, enabling modeling of microenvironment states. The dataset exhibits severe class imbalance in disease status, with approximately 80% of non-cancer cells originating from tumor tissue and 20% from healthy controls.

III. Methodology

The classification framework employs a two-layer hierarchical AdaBoost pipeline with distinct modeling strategies for each task.

A. Layer 1: Cell Type Classification

Layer 1 classifies cells into three types (Cancer, T_Cell, Fibroblast) using five diverse base learners combined through majority voting. The base learners include:

- **Decision Stump:** DecisionTreeClassifier with max_depth = 1 (50 estimators, learning_rate = 0.5)
- **Weak Tree:** DecisionTreeClassifier with max_depth = 3 (50 estimators, learning_rate = 0.5)
- **Logistic Regression:** max_iter = 600 (50 estimators, learning_rate = 0.5)
- **Naive Bayes:** GaussianNB (50 estimators, learning_rate = 0.5)
- **Linear SVM:** LinearSVC with class_weight = "balanced" (20 estimators, learning_rate = 0.3)

For each base learner, 5-fold stratified cross-validation generates out-of-fold predictions on the training set, while models trained on the full training set produce predictions on the test set. Logistic Regression, Naive Bayes, and Linear SVM use StandardScaler-normalized features, while tree-based models operate on raw feature values. The final Layer 1 prediction for each cell is determined by majority vote across all five base learners.

B. Layer 2: Disease Status Classification

Layer 2 predicts disease status (Tumor vs. Healthy_Control) exclusively for non-cancer cells identified in Layer 1. Cancer cells are automatically assigned "Tumor" status based on biological ground truth. The Layer 2 model uses augmented feature vectors created by concatenating:

- Original scaled features (7 features: Gene_A through Gene_E, Pathway_Score, UMAP_1)
- Label-encoded predictions from all five Layer 1 classifiers (5 meta-features)
- Label-encoded majority vote from Layer 1 (1 meta-feature)

This creates a 20-dimensional feature space (7 original + 6 encoded categorical + 6 duplicates with "_enc" suffix) for disease status prediction.

For each of the five base learner types, AdaBoost classifiers are trained on the non-cancer subset using the augmented features. AdaBoost configurations are:

- **Decision Stump, Weak Tree, Logistic Regression, Naive Bayes:** 50 estimators with learning_rate = 0.5
- **Linear SVM:** 20 estimators with learning_rate = 0.3

The pipeline generates predictions from all five AdaBoost models and combines them through majority voting to produce the final disease status for each non-cancer cell. This approach primarily benefits from AdaBoost's ability to focus on difficult-to-classify examples while also taking advantage of the diversity from the multiple weak learners.

C. Evaluation Strategy

We have the dataset undergo a single train-test split with 38% held out for testing, stratified by cell type to maintain class distribution. Layer 1 predictions will use cross-validation for training meta-features but are evaluated on the held-out test set. Layer 2 models are then trained only on non-cancer cells from the training set and also evaluated on non-cancer cells from the test set as identified by Layer 1's majority vote.

IV. Results

A. Layer 1: Cell Type Classification

Layer 1 achieves a 100% accuracy on cell type classification which was expected, especially because the gene expression markers (oncogene, immune, and stromal markers) provide a clear distinction between cell types. Even though high performance means that our approach is effective, it also raised some concerns that the dataset we are testing is too simple for our classification approach.

B. Layer 2: Disease Status Classification

On the other hand, classification of disease status for non-cancer cells was much more challenging to do. Individual AdaBoost classifiers trained on each base learner type provided varied performances, but we notice that simpler "weak learners" were able to outperform the more complex models when combined together with the MajorityVote approach.

TABLE I
Accuracy Summary for Layer 1 and 2 Classifications

Layer	Model	Accuracy	Accuracy_NonCancer
Layer-1	MajorityVote	1.000000	NaN
Layer-2	Decision Stump	0.873684	0.810526
Layer-2	Weak Tree	0.873684	0.810526
Layer-2	Logistic Regression	0.875439	0.813158
Layer-2	Naive Bayes	0.875439	0.813158
Layer-2	Linear SVM	0.715789	0.573684
Layer-2	**MajorityVote**	**0.874561**	**0.811842**

Layer 2 MajorityVote combines the predictions across all five AdaBoost models for the final disease status classification. This highlighted an interesting hierarchical difficulty (strong cell type classification but challenging disease status prediction) that reflects the complexity of human biology, and we concluded that identifying cell types based on marker genes is much more straightforward than characterizing their functional states within TMEs.

C. Key Insights

The results reveal several key insights:

- **Weak learners excel:** Decision Stump (max_depth = 1) and Weak Tree (max_depth = 3) achieve approximately 87.4% overall accuracy and 81.1% accuracy on non-cancer cells specifically.
- **Complex models underperform:** Linear SVM, despite being a more sophisticated model, achieves only 71.6% overall accuracy and 57.4% on non-cancer cells, substantially worse than simpler tree-based approaches.
- **Majority vote provides stability:** The final ensemble majority vote achieves 87.5% overall accuracy and 81.4% on non-cancer cells, demonstrating that combining diverse models produces robust predictions even when individual classifiers disagree.

AdaBoost is designed to work optimally with weak learners, so complex models like SVM may already be too "strong"

or "overkill" individually and won't benefit as much from AdaBoost's iterative reweighting strategy. The results align with this foundation.

V. Discussion

A. Layer 1: Dataset Simplicity and Model Selection

In our first layer, we concluded with perfect results in distinguishing the cell types of each entry. We used AdaBoosting, but one could think that this is overcomplicated. It seems that our dataset, at least for this first layer, wasn't complex enough to truly test our models. The gene expression markers provide such clear separability that even simple classifiers can distinguish Cancer, T_Cell, and Fibroblast populations with high confidence.

If we had more time, we would search for a dataset that could test this theory out more, perhaps one with overlapping marker expression or transitional cell states that would genuinely challenge the classification pipeline. After also reflecting on other groups' presentations, it would have been insightful to have a control group of our classifiers to test if the AdaBoosting was optimizing our accuracy, or if the base learners alone could achieve similar results without the boosting framework.

Although the data was simple, not every classifier with AdaBoosting was 100% correct on every individual prediction. Our system of majority vote did allow for consistency with our results and allowed us to use those results for the next step of parsing the cancer cells with little concern. This consensus mechanism proved valuable: when individual models occasionally misclassified edge cases, the majority vote corrected these errors and provided reliable cell type labels for downstream analysis.

B. Layer 2: Addressing Severe Class Imbalance

Our second layer allowed for much more interesting results. One critical issue we ran into is how the dataset is structured. With the cancer cells removed and automatically assigned "Tumor" status, the disease status of tumor-associated cells made up approximately 76% of the remaining data, leaving only 24% as healthy controls. This introduced a significant problem where our classifiers could simply predict "Tumor" for every non-cancer cell and achieve a decent accuracy of 76% without learning any meaningful patterns. Obviously, this is not what we want out of our model since we could write a trivial program that does that without any machine learning knowledge.

To fix this, we increased the test set percentage to use 38% of our data instead of the typical 20% (equivalently, reducing training data from 80% to 62%). This adjustment improved model stability and generalization. However, this does come with some tradeoffs. We still have 62% of the data for training, but with a dataset size of only 3,000 cells, we are working with a relatively limited training sample. This could lead to situations where our results are influenced by the specific train-test split, but we found our results to be pretty stable from run to run, suggesting the models successfully learned generalizable patterns rather than memorizing training examples.

One crucial lesson we learned using AdaBoosting was that it performs better with weaker classifiers, as the algorithm is designed to combine them into a stronger overall model. By focusing on misclassified examples and iteratively reweighting training samples, AdaBoost uses these errors to iteratively improve its performance. This is why, in our results, we can see that a decision tree that has a max depth of one (Decision Stump) can significantly outperform a much stronger and more complex model like a Support Vector Machine. The simple decision stump allows AdaBoost's sequential learning process to build complexity gradually, while the SVM's sophisticated decision boundary may already be "too strong" and doesn't benefit as much from the boosting iterations.

Our model produces total accuracies of around 87.5% on average, with approximately 81% accuracy specifically on the challenging non-cancer disease status prediction task. Although we could desire more from this model, our dataset size limits us. If we had a larger dataset with tens of thousands of rows, we believe that our results would be more accurate and that the models would better capture the subtle differences between tumor-associated and healthy immune/stromal cells. The current performance represents a reasonable balance given the class imbalance and limited sample size, and substantially exceeds the 76% baseline of naive majority-class prediction.

VI. Limitations

There are definitely limitations to the scope of our project. The simulated data-based approach we used likely does not have the complexity of real TMEs, especially since we determined earlier that our dataset may have been too simple for this project. Though it seems that training on one of the test datasets with one train-test split confirms consistent results, k-fold cross-validation should be employed for better estimation of the performance. We only measured 7 features in our project, but real scRNA-seq data is composed of thousands of different genes. We would probably use dimensionality reduction or optimal feature selection to help out with the computational labor if we wanted to make this project more accurate to real TME studies.

VII. Conclusion

A. Lessons Learned

1) **AdaBoost classifiers perform better with more complex datasets:** While detecting Cancer cells is easy (100% accuracy), differentiating healthy vs tumor-associated immune cells remains difficult as the biological signal is much weaker. This demonstrates the importance of matching model complexity to task difficulty.
2) **Fine-tuning data splits can be essential for increasing accuracy:** Increasing the test set size from 20% to 38% significantly improved our model's stability and generalization accuracy. This adjustment helped address class imbalance concerns and provided more robust performance estimates.

3) **Complex models do not always mean better results:** Complex models like Linear SVM failed to boost effectively while simple Decision Trees with max depth 1 and 3 thrived. AdaBoost's iterative learning process works optimally with weak learners that leave room for improvement, rather than already-sophisticated models.

To review, our project successfully deploys a two-layer hierarchical AdaBoost classification pipeline for scRNA-seq cell classification that achieves perfect cell type identification (100% accuracy) and strong disease status prediction (87.5% overall, 81.4% on non-cancer cells specifically). The value of ensemble diversity (MajorityVote) as well as using a hierarchical model for multi-task problems are highlighted especially. We also conclude that using weak learners with AdaBoost is more beneficial than individually complex models (such as SVM).

Our project architecture is particularly well-suited for imbalanced, multi-dimensional classification problems in computational biology where different tasks exhibit varying levels of difficulty. Layer 1's perfect accuracy, while validating our approach, suggests future work should test the pipeline on more challenging datasets. Layer 2's strong but imperfect performance reflects the genuine biological challenge of distinguishing tumor-associated from healthy cells.

REFERENCES

[1] G. Chen et al., "Single-cell RNA-seq technologies and related computational data analysis," *Frontiers in Genetics*, vol. 10, p. 317, 2019.

[2] D. Lähnemann et al., "Eleven grand challenges in single-cell data science," *Genome Biology*, vol. 21, no. 1, pp. 1-35, 2020.

[3] H. Chen et al., "Assessment of computational methods for the analysis of single-cell ATAC-seq data," *Genome Biology*, vol. 24, no. 1, p. 19, 2023.

[4] C. E. Meacham and S. J. Morrison, "Tumour heterogeneity and cancer cell plasticity," *Nature*, vol. 501, no. 7467, pp. 328-337, 2013.

[5] S. Aibar et al., "SCENIC: single-cell regulatory network inference and clustering," *Nature Methods*, vol. 14, no. 11, pp. 1083-1086, 2017.