Model Behavior and Predictive Stability Under Severe Class Imbalance in High-Dimensional Classification

Linxi Li LIL638@USASK.CA
University of Saskatchewan, Saskatoon, Canada

Li Xing LI.XING@USASK.CA
University of Saskatchewan, Saskatoon, Canada

Abstract

Class imbalance remains a major challenge in high-dimensional biological classification problems, particularly in single-cell RNA sequencing (scRNA-seq), where rare cell populations are often underrepresented. In this study, we evaluate the behavior of four classification models, including Elastic Net, Random Forest, Extreme Gradient Boosting (XGBoost), and multilayer perceptron (MLP), under varying imbalance and sample size conditions using a two-factor simulation framework derived from the Human Lung Cell Atlas (HLCA). Simulation settings vary across five minority class proportions and four sample size levels, with repeated train-test splits used to evaluate predictive performance. Under severe imbalance conditions, substantial differences in minority class recovery and predictive stability are observed across model architectures. Elastic Net shows unstable minority class recovery under extreme imbalance, including non-monotonic recall behavior at intermediate imbalance levels. In contrast, MLP demonstrates the fastest recovery as sample size and minority class proportion increase, while XGBoost maintains comparatively stable performance across imbalance regimes. Random Forest shows more gradual improvement as sample size increases. These results suggest that class imbalance interacts differently across model architectures and substantially influences predictive stability and minority class recovery in high-dimensional classification settings.

Keywords: class imbalance, high dimensional classification, scRNA-seq, imbalanced learning, deep learning, machine learning, simulation study

1. Introduction

High dimensional classification problems are common in scRNA-seq, where each cell is represented by the expression levels of thousands of genes [10, 12]. One major task in scRNA-seq analysis is cell type annotation, which assigns cells to predefined biological categories based on gene expression profiles [8]. Supervised learning models are widely used for this task because they can learn classification patterns directly from annotated reference datasets [1].

Despite the success of supervised learning approaches, class imbalance remains a major challenge in high dimensional biological classification [6]. In scRNA-seq datasets, abundant cell populations are often much more highly represented than rare cell types, which may contain only a small number of observations [5, 13]. Under these conditions, classification models may prioritize majority class performance, leading to poor minority class detection and unstable prediction behavior [3, 7].

Previous studies have explored imbalance handling strategies including resampling methods, class weighting, and modified loss functions [2, 4, 9]. However, less attention has been given to how different model architectures behave under varying imbalance conditions.

In this study, we evaluate four supervised classification models under controlled imbalance settings derived from the HLCA [11]. A two factor simulation framework varying both minority class proportion and total sample size is used to evaluate model performance across recall, F1 score, AUC, and balanced accuracy. This study emphasizes predictive stability and minority class recovery behavior across repeated experiments under severe imbalance conditions.

2. Experimental Design and Methods

2.1. Data Source

The data used in this study are derived from the HLCA, a large scale reference atlas containing annotated human lung single cell transcriptomic profiles [11]. To reduce dataset complexity, we select a subset of epithelial cell types from the HLCA. The resulting dataset contains 23,127 cells across nine epithelial cell types.

Quality control filtering and log normalization are performed following standard scRNA-seq preprocessing procedures. Then, the top 5,000 highly variable genes are retained as input features for all simulation experiments.

2.2. Two Factor Simulation Design

To construct a binary classification task, pulmonary alveolar type 1 cells are defined as the minority class (Class 1), while all remaining epithelial cell types are grouped as the majority class (Class 0).

To evaluate the effects of class imbalance and sample size, we design a two factor simulation framework varying both minority class proportion and total sample size. The Class 1 proportion is defined as

$$p \in \{0.01,\ 0.05,\ 0.10,\ 0.20,\ 0.30\},$$

and the total sample size is defined as

$$n \in \{500,\ 2000,\ 7000,\ 10000\}.$$

These factors produce 20 experimental settings. For each setting, cells are randomly sampled from the processed dataset according to the predefined class proportion and sample size. The sampled data are then divided into 80% training data and 20% testing data using stratified splitting to preserve the predefined class proportions in both datasets.

Within each repetition, hyperparameter tuning is performed using five fold cross validation on the training data only. The final model is then evaluated on the independent 20% testing data. This entire sampling, splitting, and evaluation procedure is independently repeated 100 times for each experimental setting.

2.3. Classification Models

Four supervised classification models are evaluated in this study, including Elastic Net, Random Forest, XGBoost, and MLP. These models represent regularized linear models, ensemble tree based methods, gradient boosting methods, and deep neural networks.

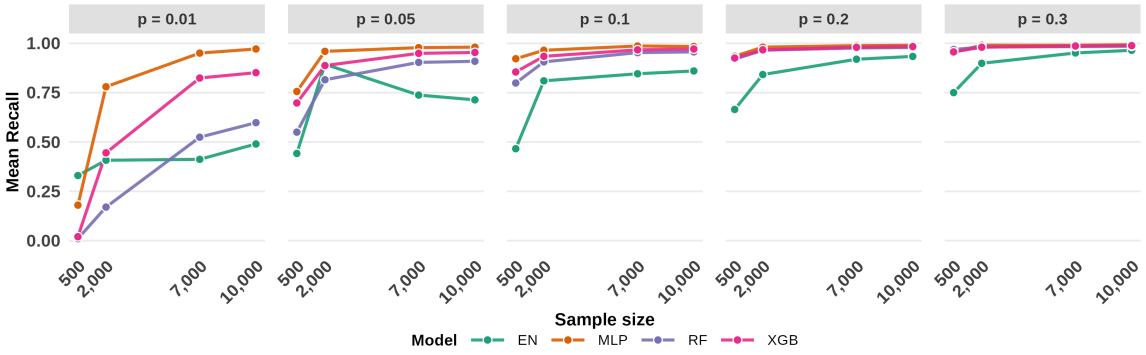

Figure 1: Line plots of mean recall across sample sizes for each class 1 proportion p. Each line represents a model-specific average recall trajectory across repeated simulations, with separate panels corresponding to different values of p.

For MLP, class weighted training is additionally applied to increase the contribution of Class 1 samples during optimization under severe imbalance conditions.

2.4. Evaluation Metrics

Model performance is evaluated on independent testing data using recall, precision, F1 score, balanced accuracy, and AUC. This study primarily focuses on recall, balanced accuracy, and performance distributions across repeated experiments. Formal metric definitions are provided in the Appendix.

3. Empirical Analysis

3.1. Effects of Class Imbalance and Sample Size

Figure 1 summarizes the average recall across different combinations of minority class proportion and sample size for all four models. Overall, recall increases as either the minority class proportion or the total sample size increases. However, different models respond differently to increasing sample size and minority class proportion.

Under severe imbalance conditions, particularly when $p = 0.01$, Elastic Net shows substantially weaker Class 1 detection than the other models. Recall remains relatively low even as sample size increases, suggesting that the number of available Class 1 samples is insufficient for stable minority class recovery under extreme imbalance.

Appendix Figure A1 further illustrates the recall distributions of Elastic Net across repeated experiments under severe imbalance conditions. In many runs, recall remains close to zero, while other runs achieve substantially higher Class 1 detection. Consistent with this pattern, Appendix Table A1 shows that severe imbalance frequently produces undefined precision and F1 scores, indicating complete failure of Class 1 detection in repeated experiments.

Interestingly, Elastic Net exhibits a distinct transition pattern around $p = 0.05$. Recall increases substantially from $n = 500$ to $n = 2000$, but declines at larger sample sizes. Unlike the other models, Elastic Net does not show consistent improvement as sample size increases under moderate

Figure 2: Heatmap of mean balanced accuracy across simulation settings. Each panel represents a model, with sample size n on the y axis and Class 1 proportion p on the x axis. Cell colors and labels indicate the average balanced accuracy across repeated simulations.

imbalance conditions. One possible explanation is related to the coefficient shrinkage mechanism of Elastic Net. As sample size increases under fixed imbalance, the optimization procedure may increasingly favor stable majority class coefficients, causing weaker minority associated signals to be shrunk toward zero and reducing Class 1 recall.

When $p \geq 0.10$, Elastic Net shows more consistent improvement as sample size increases, suggesting that the number of Class 1 samples becomes sufficient for more stable coefficient estimation and minority class recovery. In contrast, MLP shows rapid improvement in recall once either sample size or Class 1 proportion increases, although performance remains unstable under the most extreme imbalance setting ($p = 0.01$, $n = 500$).

One possible explanation for the strong performance of MLP is the use of class weighted training, which assigns greater importance to Class 1 samples during model training under severe imbalance conditions. However, weighted training alone does not fully prevent complete failure of Class 1 detection under the most extreme imbalance setting ($p = 0.01$, $n = 500$), where undefined precision and F1 scores remain common (Appendix Table A1). However, once either sample size or Class 1 proportion increases, undefined precision and F1 scores become much less common, suggesting that weighted training becomes more effective when more Class 1 samples are available. Emphasizing Class 1 samples during training also does not substantially increase predictive instability, suggesting that the minority class signal remains sufficiently informative for stable learning under the evaluated simulation settings.

XGBoost also shows strong and relatively stable recall across most imbalance regimes, while Random Forest demonstrates more gradual improvement as sample size increases. As the minority class proportion increases, performance differences between models become substantially smaller, with all models approaching high recall under moderate imbalance settings. These results suggest that model specific differences become most pronounced when the number of available Class 1 samples is extremely limited.

3.2. Balanced Accuracy Across Experimental Regimes

Figure 2 summarizes the mean balanced accuracy across all combinations of minority class proportion and sample size for the four classification models.

Across all models, balanced accuracy generally increases as either the minority class proportion or sample size increases. Under severe imbalance conditions, particularly when $p = 0.01$, Elastic

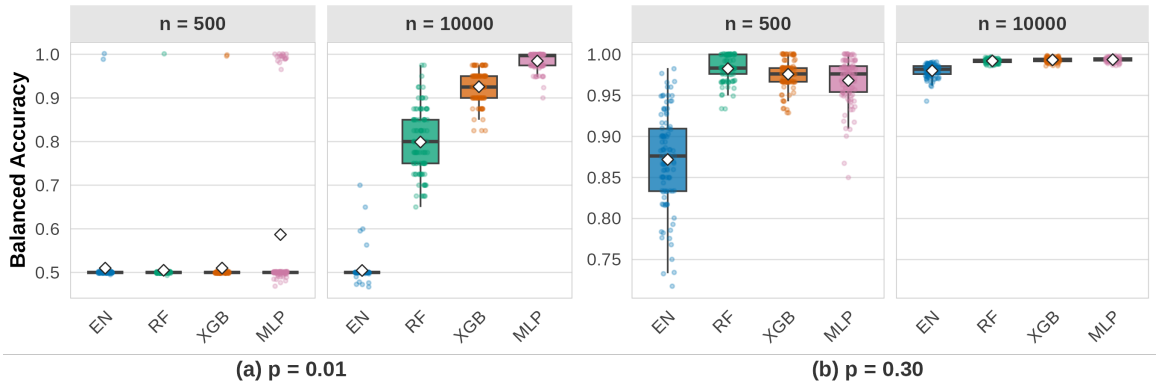

Figure 3: Boxplots of balanced accuracy across four representative simulation settings corresponding to combinations of severe versus moderate imbalance and small versus large sample sizes. Each panel shows the distribution of balanced accuracy across repeated simulations for a given model.

Net frequently produces balanced accuracy values close to 0.5, indicating near random Class 1 detection. This result is consistent with the recall instability observed previously.

In contrast, MLP demonstrates the fastest recovery in balanced accuracy as either sample size or Class 1 proportion increases. Except for the most extreme imbalance setting ($p = 0.01$, $n = 500$), MLP rapidly achieves high balanced accuracy across neighboring simulation regimes. XGBoost also maintains comparatively stable performance across most settings, while Random Forest demonstrates more gradual improvement as sample size increases.

Figure 3 further illustrates the distribution of balanced accuracy across four representative combinations of severe versus moderate imbalance and small versus large sample size settings.

Under extreme imbalance conditions ($p = 0.01$, $n = 500$), all models produce balanced accuracy values concentrated near 0.5, indicating consistently poor Class 1 recovery rather than stable predictive performance. This result shows that low variability does not always indicate good model performance under severe imbalance, since models may repeatedly fail to detect the minority class across experiments.

As sample size increases under severe imbalance ($p = 0.01$), the performance differences between models become much more visible. Elastic Net remains centered near 0.5 even at $n = 10000$, suggesting that increasing sample size alone provides limited improvement under extreme imbalance conditions. In contrast, Random Forest, XGBoost, and MLP show progressively higher balanced accuracy distributions as sample size increases, although Random Forest remains comparatively more variable across repeated experiments.

Under moderate imbalance conditions ($p = 0.3$), all models show substantial improvement in balanced accuracy relative to severe imbalance settings. However, model specific differences remain visible at smaller sample sizes ($n = 500$), where Elastic Net continues to exhibit lower balanced accuracy and greater variability than the other models. These differences become substantially smaller at larger sample sizes ($n = 10000$), where all models achieve highly concentrated balanced accuracy distributions near 1.0.

References

[1] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20(1):194, 2019.

[2] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[4] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[5] Dominic Grün and Alexander van Oudenaarden. Design and analysis of single-cell sequencing experiments. *Cell*, 163(4):799–810, 2015.

[6] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[7] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

[8] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[10] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[11] Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C Strobl, Tessa E Gillett, Luke Zappia, Elo Madissoon, Nikolay S Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, et al. An integrated cell atlas of the lung in health and disease. *Nature medicine*, 29(6):1563–1577, 2023.

[12] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

[13] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25 (10):1491–1498, 2015.

Appendix A. Supplementary Results

Additional metric definitions and supplementary analyses are provided in this appendix.

A.1. Evaluation Metric Definitions

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

F1 score:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Balanced accuracy:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

A.2. Undefined Precision and F1 Statistics

Table A1: Counts of undefined precision and F1 scores across repeated simulations under different imbalance settings. Undefined values occur when models produce no positive predictions, indicating complete failure of Class 1 detection under severe imbalance conditions.

Model	p	n	Precision NA	F1 NA	Precision NA Rate	F1 NA Rate
EN	0.01	500	66	67	0.66	0.67
EN	0.01	2000	52	53	0.52	0.53
EN	0.01	7000	39	53	0.39	0.53
EN	0.01	10000	36	47	0.36	0.47
EN	0.05	500	43	44	0.43	0.44
EN	0.05	2000	2	4	0.02	0.04
EN	0.05	7000	6	14	0.06	0.14
EN	0.05	10000	9	17	0.09	0.17
MLP	0.01	500	68	82	0.68	0.82
MLP	0.01	2000	0	1	0.00	0.01
MLP	0.01	7000	0	0	0.00	0.00
MLP	0.01	10000	0	0	0.00	0.00
MLP	0.05	500	2	2	0.02	0.02
RF	0.01	500	98	99	0.98	0.99
RF	0.01	2000	52	52	0.52	0.52
RF	0.05	500	4	4	0.04	0.04
XGB	0.01	500	98	98	0.98	0.98
XGB	0.01	2000	11	11	0.11	0.11
XGB	0.05	500	3	3	0.03	0.03

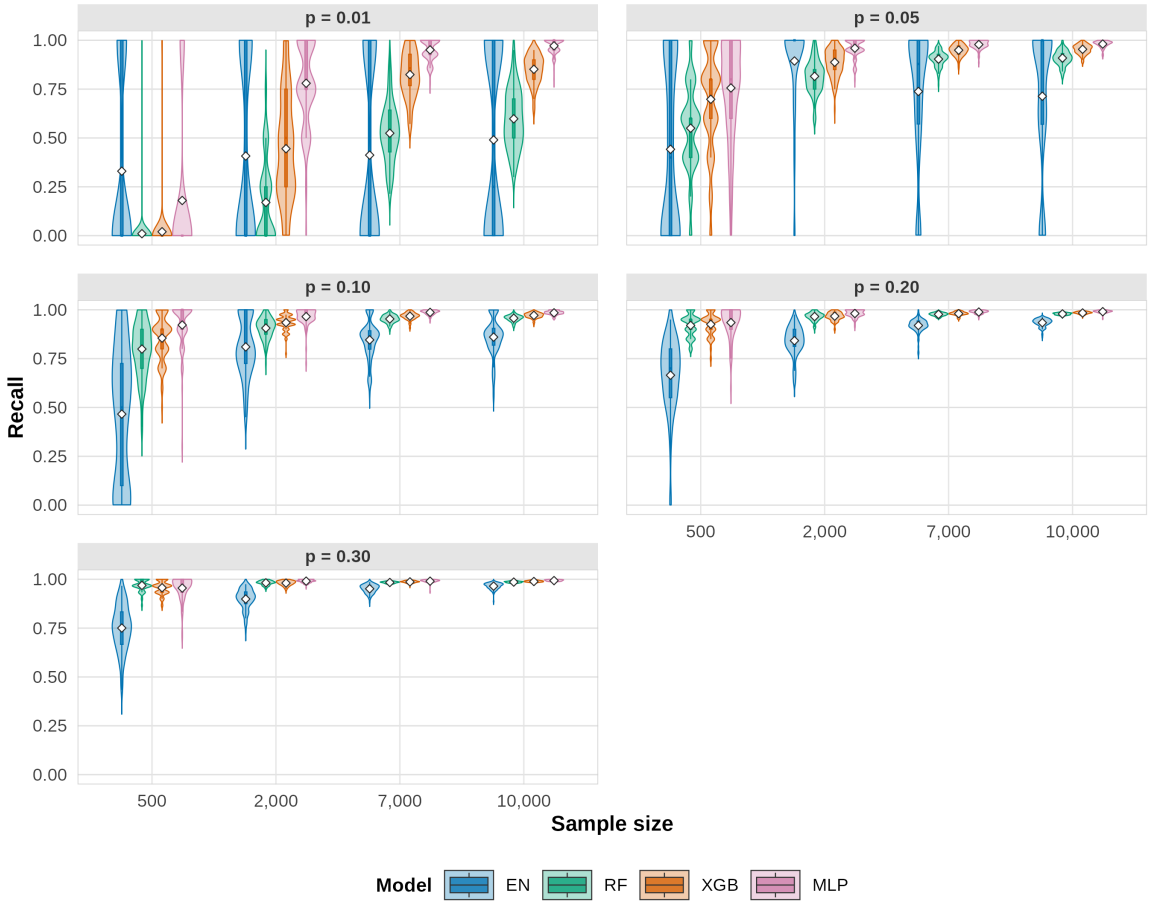

Figure A1: Violin plot distributions of recall across repeated simulations under different Class 1 proportions and sample sizes. Each panel represents one Class 1 proportion setting. The x axis represents sample size, and the y axis represents recall.