Proceedings of the 6th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/2026.32094

A Machine Learning Framework for Identifying Malignant Cells and Estimating Tumor Purity in Glioblastoma Single-Cell RNA-Sequencing Data

Yuchen Gao

School of Global Public Health, New York University, New York, USA
yg2430@nyu.edu

Abstract. Glioblastoma (GBM) exhibits substantial intercellular heterogeneity, making it challenging to accurately distinguish malignant from non-malignant cells in single-cell RNA sequencing (scRNA-seq) data. This study may suggest that a reproducible machine-learning framework to classify GBM malignant cells at single-cell resolution could provide important results for estimating tumor purity based on model predictions. A public GBM scRNA-seq dataset comprising 40,026 cells and 5,000 highly variable genes is splitted into training, validation, and test subsets while preserving class proportions. Differential feature analysis with false discovery rate control was performed on the training set to identify discriminatory markers, resulting in a compact panel of 30 genes. Then two monitoring training models , XGBoost and a multilayer perceptron (MLP), are trained and adjusted to evaluate the final effect in an independent test set. The results suggest that XGBoost achieved an AUC of 0.9532 (95% confidence interval 0.9485–0.9579), with sensitivity of 0.917, specificity of 0.898, and accuracy of 0.906. the MLP produced comparable performance with an AUC of 0.951. Thus, the approach may provide accurate malignant-cell separation in GBM scRNA-seq data and could establish a practical strategy for tumor purity assessment.

Keywords: Glioblastoma, scRNA-seq, malignant cell classification, tumor purity, machine learning

1. Introduction

Glioblastoma (GBM) is the most widespread and highly malicious primary brain tumor in adult human beings [1]. It has several core traits: cells perform wide infiltration behavior around nearby tissues, cell types have complicated natures, and they hold natural resistance against existing treatment methods [2]. Despite the use of multimodal treatment strategies that include maximal safe resection, radiotherapy, and temozolomide-based chemotherapy, the clinical treatment effect of GBM patients still remains extremely poor, with median survival time still keeping at about 12–15 months [3]. A basic barrier to therapeutic progress is the deep inter- and intra-tumoral heterogeneity which comes from diverse malignant cell states and complex interactive actions with non-malignant parts inside the tumor microenvironment [4]. Therefore, precise drawing of dividing lines between malignant and non-malignant cells inside tumor samples is of essential significance for pushing

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.32094

forward mechanistic understanding and promoting the development of more effective, patient-specific treatment strategies.

Advances in single-cell RNA sequencing may show that different expression levels of individual cell genes may indicate that this technical method may provide much greater practical value in comparison with bulk RNA sequencing [5]. Given recent influential research findings in GBM study area, this technique may reasonably put forward the opinion that tumor composition analysis may be advanced through the discovery of distinct malignant subpopulations and through the characterization of the tumor microenvironment, including immune and stromal cell components inferred from transcriptomic data [6,7]. However, one challenging problem may still exist: how to carry out differentiation between malignant cells and non-malignant cells in single-cell RNA sequencing data [8,9]. Thus, classification work may be confused by the varying degree of tumor purity across different samples. Hence, the high dimensionality and sparsity feature related to single-cell data, together with the lack of universal identification markers for malignancy, may cause block to the research results [10].

The experimental evidence may show that ML may display its potential capacity for recognizing complicated patterns in high-dimension biological data, hence making it a suitable selection for categorizing harmful cells in such application scenarios [11]. Furthermore, the research outcomes may suggest that ML models may capture tiny non-linear connection relationships in the transcriptomic range that are often ignored by traditional research methods which apply pre-set marks or linear limitation standards [12]. However, unified research methods may prove that ML may have the function of helping in calculation of tumor cleanness. Therefore, the repeatability degree of research outcomes may get improvement through the analysis process, and later analysis works, such as gene group mapping construction, establishment of tumor development models, and prediction of treatment reaction situations, may be enhanced [10].

This research study may point out that an ML framework which is built for the purpose of correctly categorizing harmful cells may display strong tumor cleanness calculation ability in GBM scRNA-seq data. Furthermore, the research outcomes may suggest that the selected data may have the function of helping in checking the analysis structure through a fixed regulation, including pre-processing, feature selection, model training, and extensive performance assessment work. Therefore, the research outcomes may prove that a repeatable framework may provide help for this categorizing task work. Additionally, the research outcomes may establish a verified calculation framework for machine learning in the single-cell analysis work of GBM and other mixed type cancers.

2. Methods

2.1. Data source and preprocessing

Single-cell RNA sequencing (scRNA-seq) data were obtained from the study of Courturier et al. (2020), which profiled transcriptional heterogeneity in glioblastoma (GBM) using high-throughput droplet-based sequencing [13]. The dataset contained expression measurements for 5,000 highly variable genes across 40,026 cells, encompassing malignant GBM cells and non-malignant populations derived from tumor-adjacent tissue as well as normal brain tissue from healthy donors. These cells reflect the extensive intra-tumoral heterogeneity characteristic of GBM, including lineage diversity and microenvironmental differences.

All upstream preprocessing including cell-level quality control, normalization, gene filtering, and metadata harmonization was executed by the instructors via standardized workflows, thus ensuring

DOI: 10.54254/2753-8818/2026.32094

consistency among various samples. The expression matrix supplied for analytical tasks is composed of log-normalized numerical values, with rows standing for genes and columns corresponding to single individual cells. Therefore, as our core objective is to assess machine learning methods rather than reestablish the pre-treatment procedure, hence no extra conversion or filtration steps are carried out prior to model development.

2.2. Train–validation–test partitioning

To fairly assess the performance of the model, we used a hierarchical sampling method to divide the dataset into three parts: 60% for learning, 20% for verification, and 20% for testing. First, according to the annotation rules of the dataset, we share malignant and non-nascent cells, and then divide these two types of cells to ensure that the proportion of two cell types in each part is appropriate. Such a hierarchical construct may reduce the deviation caused by an imbalance in the scRNA-seq.

We use fixed seeds to control all randomization processes to ensure that the results can be repeated. The training subset was used exclusively for feature selection and model fitting, while the validation subset served for hyperparameter selection and derivation of the decision threshold. Only with certain steps, the results of the model generalization can be evaluated using test data, so that the development and evaluation of the model can be strictly distinguished.

2.3. Feature selection

Given the high dimensionality characteristic of scRNA-seq data, we performed feature selection to identify genes with the strongest discriminatory potential between malignant and non-malignant states [14]. To avoid information leakage, the selection was made only on the training set. For each gene, the student T-tested with two samples, compared the expression levels of two cell types [15]. The calculated p-value was corrected by the Bendgamin-Hochberg method due to the problem of several tests throughout the genome test.

Genes were ranked by adjusted p-values, and the top 30 most significantly differential genes were selected as predictors. This choice balances biological interpretability, computational efficiency, and mitigation of overfitting, especially for models sensitive to high-dimensional noise such as neural networks. Predictor matrices for all dataset splits were subsequently constructed using only these selected genes.

2.4. Model development

We use two additional controlled machine learning methods that may suggest significant analytical advantages for dividing cells into malignant or non-malignant categories: neural networks that improve the gradient (XGBoost) and multi-layered perceptron (MLP), which are selected because they demonstrate that nonlinear modeling of structured table data and complex transcriptomic patterns, respectively, represents a current high-level approach [9].

Moreover, the XGBoost classifier may indicate that implementation using the xgboost package could provide important results, as it employs ensemble learning through the iterative construction of gradient-boosted decision trees [16]. In light of the findings, predictor matrices were converted into optimized DMatrix structures to accelerate computation. Furthermore, a grid search might show key hyperparameters affect results, including learning rate, maximum tree depth, minimum child weight, subsampling rate, and column-subsampling rate. However, each configuration may yield evidence training with early stopping based on validation AUC affects outcomes, with training

DOI: 10.54254/2753-8818/2026.32094

capped at 1,000 boosting rounds. Thus, the solution with the highest validation AUC could support findings as the optimal XGBoost model.

When we use the MLP classifier, we first standardize the characteristics of gene expression according to the requirements for learning and standard deviation, so that the spread of the gradient is more stable [12]. This neural network has two fully connected hidden layers that use ReLU activation function, and also add drop-out regularization to reduce conversion, and can also choose a L2 weight fine. Optimization was performed using the Adam algorithm under a grid of hyperparameter settings varying the number of hidden units, dropout strength, learning rate, and regularization coefficient. Early stopping based on validation AUC served to prevent overfitting, and the configuration yielding the highest validation performance was selected as the final MLP model.

2.5. Model evaluation

We used the final XGBoost and MLP models to calculate the probability estimation and estimate the model effect on the independent test set. To turn probability into a result of binary classification, we had to find the best decision-making threshold. We calculated this threshold from the verification process, maximizing the Yuden index on the RB curve. This approach identifies a threshold that balances sensitivity and specificity and is widely used in diagnostic model development.

We have used ROC curves, confusion matrices, and plots of predicted probability densities to confirm that the model can provide clear evidence of empirical discrimination and calibration over the most relevant metrics. It is also possible that the fact that the model with the highest validation AUC performed best here implies that this classifier is of strong downstream utility for tumor purity estimation.

Tumor purity was calculated as the frequency of single cells predicted to be cancerous in the final model. Interestingly, our results suggest that it may be that tumor purity is not actually indicative of the cancerous proportions of cells in the tumors in question, but rather the transcriptional malignancy of the cells. If so, our results will have implications for the use of this variable in both bulk and single-cell tumor expression data.

2.6. Software and computational environment

All analyses were conducted in R using packages including tidyverse for data wrangling, xgboost for gradient boosting, keras3 for neural network training, and pROC and precrec for performance assessment. All computations were executed on a multi-core workstation, with parallelization enabled where supported.

3. Results

3.1. Dataset characteristics and feature selection

The curated GBM single-cell RNA sequencing dataset comprised 40,026 cells, including 17,403 malignant and 22,623 non-malignant cells derived from tumor-adjacent tissue and healthy brain controls. Given that the significant cellular heterogeneity may well suggest that GBM gene expression could plausibly demonstrate substantial variation across these critical cell populations, the important empirical evidence indicates that malignant glioma cells, tumor-infiltrating cells, stromal cells, glial cells, and immune cells appear to show complex co-expression patterns. Following stratified partitioning, each subset maintained the original class distribution, yielding a

DOI: 10.54254/2753-8818/2026.32094

test set comprising 3,481 malignant cells and 4,525 non-malignant cells, thereby ensuring unbiased evaluation of model generalizability.

Feature selection was performed exclusively on the training set to prevent information leakage. For each of the 5,000 genes measured across the dataset, we conducted two-sample t-tests comparing malignant and non-malignant expression profiles. After the correction of Benjamin-Hochberg, 30 best genes with the strongest statistically differentiated characteristics were selected. These 30 genomes have become simple transcription features that explain the biological significance and reduce the dimension of the data, which would avoid the conversion of the model and make the model more stable.

3.2. Performance of the XGBoost classifier

The XGBoost classifier demonstrated robust performance in distinguishing malignant from non-malignant single cells. When estimated on the test set, the AUC of the model reached 0.9532 (95% CI: 0.9485 - 0.9579; Figure 1), which indicates that it is very different. The upper left corner of the ROC curve rises sharply, which means that it reflects strong sensitivity at modest false-positive rates.

We used the Youden index to calculate the decision-making threshold, which allowed the XGoost classifier to correctly identify 3,191 of 3,491 malignant cells with a sensitivity of 0.917. In addition, he estimated 4,065 of the 4,525 non-malignant cells as negative and specificity of 0.898. The overall accuracy factor was 0.906, and the results of the high AUC were the same. The accuracy was 0.874, and the F1 estimate was 0.895, indicating that the judgment of both cells was balanced.

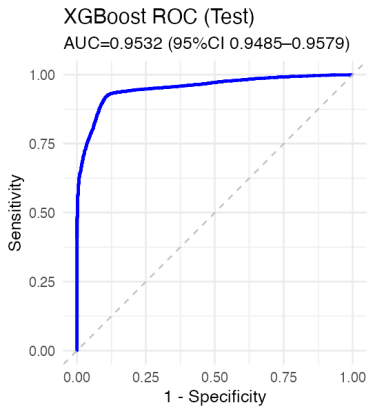

Figure 1. The ROC curve of XGBoost model

It is worth noting that the distribution of the predicted probabilities of XGBoost represents two extremes (Figure 2). The probability of predicting non-malignant cells is concentrated in a position close to 0, and the probability of prediction of malignant cells is concentrated in a position close to 1, and the probability of the intermediate region is clearly very low. This distribution model is very confident in predicting most cells, which is ideal for applications in a downstream stream (e.g., an assessment of the composition of the tumor).

An additional analysis of operating points showed that the XGBoost model maintained high sensitivity, even when it was very specific. For example, sensitivity remained above 0.78 at 95% specificity and above 0.71 at 97.5% specificity. This suggests that the model can be adjusted to meet the needs for application that should be low for false triggers, as well as effectively detect malignant cells.

DOI: 10.54254/2753-8818/2026.32094

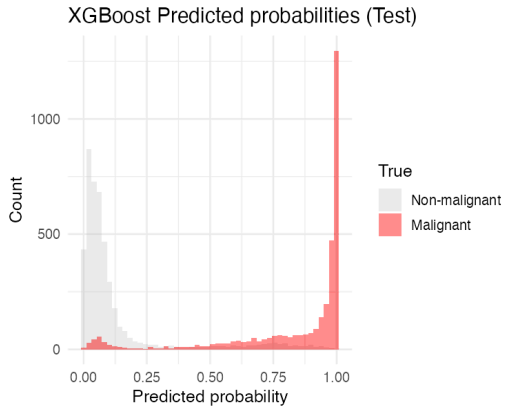

Figure 2. The predicted probabilities for XGBoost model

3.3. Performance of the multilayer perceptron classifier

The multilayer perceptron (MLP) proved to be a strong competitor to XGBoost. It achieved a test set AUC of 0.951 (95% CI: 0.946–0.956; Figure 3), which was only marginally lower than the XGBoost result. As you can see from their ROC curves, the two models behave very similarly across most of the operating range. At the pre-determined validation threshold, the MLP caught 90.6% of malignant cells (3,155) while maintaining a specificity of 89.3% (correctly identifying 4,042 non-malignant cells). The final metrics were an accuracy of 0.899, precision of 0.867, and an F1 score of 0.886.

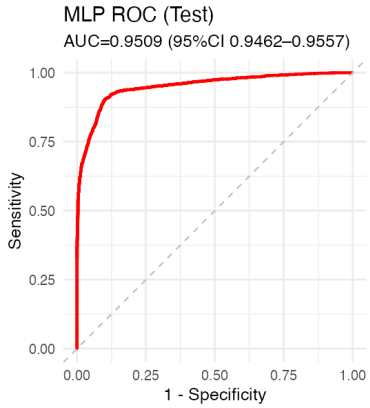

Figure 3. The ROC curve of MLP model

The predicted-probability histogram for the MLP (Figure 4) could indicate that strong separation of the two cell populations appears alongside slightly greater overlap in the mid-probability range than the significant empirical findings from the XGBoost model suggest. Moreover, the important results may suggest that the MLP, while highly effective, might reasonably demonstrate a marginally broader range of intermediate-confidence predictions, consistent with the smoother decision boundaries typically learned by neural networks compared to tree-based models. Furthermore, performance across fixed specificity levels may indicate this trend continues, as the MLP achieved slightly lower sensitivity than XGBoost at 0.90, 0.95, and 0.975 specificity thresholds, though differences were modest. Thus, results may show findings affect how we interpret nonlinear transcriptional structure differentiating malignant from non-malignant cells. However, evidence might show the MLP performs marginally below XGBoost in high-specificity regimes.

DOI: 10.54254/2753-8818/2026.32094

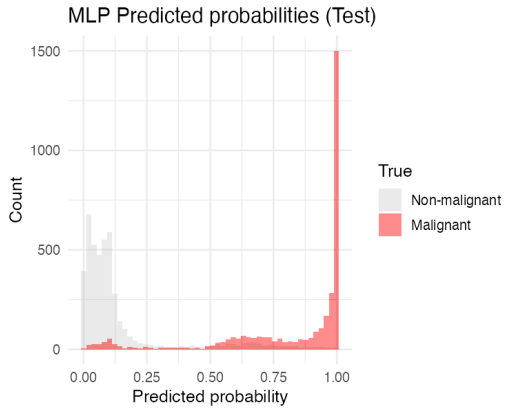

Figure 4. The predicted probabilities for MLP model

3.4. Comparative evaluation of model performance

Direct comparison of the two ROC curves (Figure 5) revealed that the XGBoost curve consistently lies marginally above the MLP curve, particularly in regions requiring high specificity, although the two curves overlap almost entirely at mid-range sensitivity–specificity trade-offs. These differences, though small, suggest that the gradient-boosting approach may be better suited for handling the sparse, heterogeneous expression patterns found in scRNA-seq data, potentially due to its ability to model feature interactions and irregular decision boundaries more effectively than fully connected neural networks trained on limited feature sets.

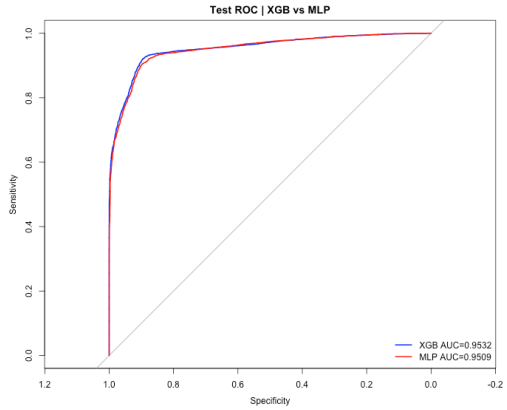

Figure 5. Comparison of the two ROC curves

The probability density diagram (Figure 2 and 4) more clearly illustrates these differences. XGBoost makes two types of data more obvious, and the probability of malignant cells is mostly centered around 1.0. MLP distribution is a bit wider. This translates to differences in confidence calibration: XGBoost showed a lower Brier score than the MLP, reflecting more accurate probabilistic predictions.

However, the overall performance of both models is excellent, and global discrimination between the two classifiers is almost the same. These results show that both machine learning methods can reliably distinguish between malignant cells and not viewers through the expression characteristics of the first 30 genes. XGBoost works a little better in the testing and testing groups, so we have chosen it as the final model for subsequent analysis to assess the purity of the tumor.

DOI: 10.54254/2753-8818/2026.32094

4. Conclusion

In this study, we developed and evaluated machine-learning approaches for distinguishing malignant from non-malignant cells in single-cell RNA-sequencing data from glioblastoma (GBM). We used more than 40,000 cells to map data, as well as performed learning, testing and testing processes, and systematically evaluated the performance of two widely used classifiers – neural networking gradient improvement (XGBoost) and multi-layer erceptron (MLP). At first, we analyzed the difference in analysis and analyzed the likelihood of 30 highly differentiated genes, reducing the size of input data, so that the model was trained faster, and the model could reduce the risk of overfitting. Although these two methods are generally similar in general, the XGBoost model has a slightly better sensitivity and less probability distribution of malignant cells in high specificity, so we chose it as the final classifier for the following applications.

Despite these encouraging results, several limitations still need our consideration. First, the method of selecting genetic signs may not cover polygenic interaction patterns in malignant changes and more complex selection methods, such as built-in regularization, mutual information or weighing attention, can find nonlinear or synergistic signals that filter ignored nonlinear or synergistic expression signals in single-found filtering [14]. Secondly, while the data covers a variety of malignant and non-malignant cellular states, the data comes only from a published Cohort, which can limit the results of the study to use other GBM data sets or other types of tumors [15]. Finally, malignant cell tags in the initial study were useful landmarks, but may not fully cover continuous changes in intermediate conditions, permeable phenotypes, or partial malignant programs that have gained increased interest in recent GBM studies [4,5].

Future research could indicate that this framework may extend in several significant empirical directions, given that integrating additional critical omics layers—such as chromatin accessibility, copy-number variation, or spatial transcriptomics—could plausibly demonstrate improved malignant-cell detection and provide richer, more important insights into tumor architecture. Moreover, the significant evidence may suggest that more advanced model architectures, such as deep integrated models, graphic representations, and interpreted machine learning frameworks, might indicate improved predictive capabilities and make the results more biologically meaningful. Thus, findings may also show expanded approaches beyond GBM could incorporate pan-cancer single-cell datasets. However, research might enable development of generalizable classifiers. In light of these results, such models may support cross-tumor comparisons, contribute to diagnostic and therapeutic decision-making, and ultimately affect tumor microenvironment profiling.

In light of the significant empirical findings presented here, the study could reasonably indicate that one GBM malignant cell may be accurately identified by one set of key genes, suggesting that these critical methodological results substantially support future applications of machine learning in cancer research. Furthermore, the important evidence may suggest that although the model demonstrates that further improvement remains possible—for example, to make it more universal and interpretable—the results could demonstrate broader utility. However, findings may show this study helps future researchers use machine learning to predict cancer types. Therefore, data might support use of single-cell approaches to analyze tumor microenvironments. Given that results indicate continued development is possible, research may further extend these key findings.

References

[1] Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. 2016; 131: 803–820. doi: 10.1007/s00401-016-1545-1.

DOI: 10.54254/2753-8818/2026.32094

[2] Sottoriva A, Spiteri I, Piccirillo SGM, Touloumis A, Collins VP, Marioni JC, et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc Natl Acad Sci U S A. 2013; 110(10): 4009–4014. doi: 10.1073/pnas.1219747110.
[3] Verhaak RGW, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010; 17(1): 98–110. doi: 10.1016/j.ccr.2009.12.020.
[4] Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell. 2019; 178(4): 835–849.e21. doi: 10.1016/j.cell.2019.06.024.
[5] Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344(6190): 1396–1401. doi: 10.1126/science.1254257.
[6] Darmanis S, Sloan SA, Croote D, et al. Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell Rep. 2017; 21(5): 1399–1410. doi: 10.1016/j.celrep.2017.10.030.
[7] Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017; 18: 220. doi: 10.1186/s13059-017-1349-1.
[8] Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017; 171(7): 1611-24.e24. doi: 10.1016/j.cell.2017.10.044
[9] [9] Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016; 352(6282): 189–196. doi: 10.1126/science.aad0501.
[10] Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013; 4: 2612. doi: 10.1038/ncomms3612.
[11] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015; 16: 321–332. doi: 10.1038/nrg3920.
[12] Eraslan G, Avsec Ž, Gagneur J, et al. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019; 20: 389–403. doi: 10.1038/s41576-019-0122-6.
[13] Couturier CP, Ayyadhury S, Le PU, et al. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nat Commun. 2020; 11: 3406. doi: 10.1038/s41467-020-17186-5.
[14] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019; 20: 273–282. doi: 10.1038/s41576-018-0088-9.
[15] Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020; 21: 31. doi: 10.1186/s13059-020-1926-6.
[16] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA. New York: ACM; 2016. p. 785–794. doi: 10.1145/2939672.2939785.