Proceedings of ICBioMed 2025 Symposium: Computational Modelling and Simulation for Biology and Medicine
DOI: 10.54254/2753-8818/2025.LD27586

Classification of Malignant and Non Malignant Cells in Glioblastoma Based on Single-Cell Rna Sequencing and Machine Learning

Xiaodi Wang[1*], Kai Fu[2]

[1]Tianjin University, Tianjin, China
[2]UCLA, Los Angeles, USA
**Corresponding Author. Email: 1853409129@qq.com*

Abstract. Glioblastoma (GBM) is a highly aggressive primary brain tumor with significant intra - and inter - tumoral heterogeneity, posing great challenges to accurate diagnosis and effective treatment. In this study, we leveraged single - cell RNA sequencing (scRNA - seq) technology and machine learning algorithms to conduct a comprehensive investigation into the classification of malignant and non - malignant cells in GBM.First, scRNA - seq data was used to profile the gene expression patterns of different cell populations within GBM tissues. Then, a series of machine learning algorithms, including random forest, were applied for cell classification. The performance of the model was evaluated using multiple metrics such as accuracy, precision, recall, and F1 - score.The results showed that the model achieved an overall accuracy of 87%. For the two classes (malignant and non - malignant cells), the precision ranged from 85% - 89% and the recall spanned 86% - 88%, with F1 - scores of 0.86 and 0.89 respectively. Both the macro - average and weighted - average F1 - scores reached 0.87, indicating a relatively balanced classification performance. However, from the confusion matrix, there were still 985 false - negative cases and 1048 false - positive cases, suggesting room for improvement in distinguishing between the two cell types.In conclusion, this study demonstrates the potential of integrating scRNA - seq and machine learning in GBM cell classification, while also identifying areas for further optimization to enhance the accuracy of distinguishing malignant and non - malignant cells, which is of great significance for improving GBM diagnosis and treatment strategies.

Keywords: Glioblastoma, Single - cell RNA Sequencing, Machine Learning, Cell Classification, Malignant and Non - malignant Cells

1. Introduction

1.1. Single-cell RNA sequencing (scRNA-seq): resolving cellular heterogeneity and its relevance to glioblastoma multiforme (GBM)

Single-cell RNA sequencing (scRNA-seq) enables gene expression analysis at individual cell resolution, offering detailed insights into cellular heterogeneity and rare cell populations. Unlike

© 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: 10.54254/2753-8818/2025.LD27586

traditional bulk RNA sequencing, scRNA-seq accurately identifies individual cells within complex tissues, particularly valuable in tumor biology where the tumor microenvironment (TME) includes diverse cell types, such as malignant cells, immune cells, and stromal cells. Glioblastoma multiforme (GBM) is the most aggressive primary brain tumor in adults, with a median survival of around 15 months despite standard treatment (surgery, radiation, and temozolomide chemotherapy) [1]. GBM's poor prognosis is largely due to its significant intratumoral and intertumoral heterogeneity, featuring multiple subclones and distinct molecular subtypes (classical, mesenchymal, proneural).

1.2. The complex Tumor Microenvironment (TME), key molecular alterations, and the need for accurate tumor cell identification in GBM

The TME in GBM is complex, involving various non-tumor cells like tumor-associated macrophages (TAMs), oligodendrocyte precursor cells (OPCs), and endothelial cells, significantly impacting tumor progression and therapy resistance. Common molecular alterations in GBM include EGFR amplification, rare IDH1/2 mutations, and PTEN deletion or mutation. Elevated transcriptional markers, such as CHI3L1 and CD44 in mesenchymal GBM, correlate with poor outcomes. Accurate tumor cell identification via scRNA-seq is essential for understanding tumor-driving mechanisms, developing precise therapeutic targets (e.g., EGFRvIII), optimizing immunotherapy strategies, and identifying rare but crucial subpopulations like glioma stem cells (GSCs).

1.3. Recent advances of scRNA-seq in GBM research and the rationale for machine learning-based malignant cell classification

Recent scRNA-seq studies have significantly advanced GBM research. Detection of copy number variations (CNVs) and mutations (e.g., IDH1 R132H) differentiates malignant from non-malignant cells. A notable 2021 Nature study revealed interactions between SPP1+ macrophages and tumor cells promoting GBM invasiveness. Utilizing machine learning to analyze scRNA-seq data from GBM patients presents a promising approach for accurately identifying tumor cells, improving diagnostic precision, discovering novel therapeutic targets, and enabling personalized treatment strategies [2]. In this manuscript, we applied machine learning classification methods to scRNA-seq data from GBM patients to systematically distinguish malignant cells from non-malignant cells. The following sections detail our methodological approaches, present our results, and discuss their implications for GBM diagnosis and targeted therapy.

2. Process and methods

2.1. Dataset description

Dataset. The dataset comprises standardized gene expression values for 5,000 genes across 49,026 single-cell samples. Samples are clearly labeled: malignant samples have column names starting with "Malignant" (marked as 1), and non-malignant samples have column names beginning with "Non_malignant" (marked as 0). Each entry represents the standardized expression level of a gene.

2.2. Analytical tools and methods

Analytical Tools and Methods. The analysis was performed using several Python libraries:

DOI: 10.54254/2753-8818/2025.LD27586

●NumPy (np): Efficient numerical computations, array, and matrix manipulations.
●Pandas (pd): Data manipulation and cleaning.
● RandomForestClassifier (scikit-learn): Random forest algorithm for ensemble-based classification.
●MLPClassifier (scikit-learn): Multi-layer perceptron neural network classifier.
●SVC (scikit-learn): Support vector machine classifier.
●XGBoost: Gradient boosting framework for classification and regression.
●train_test_split (scikit-learn): Divides data randomly into training and testing subsets.
● Metrics for model evaluation including classification_report, accuracy_score, roc_auc_score, confusion_matrix (scikit-learn).
●Ranksums (SciPy): Wilcoxon rank-sum test, a non-parametric statistical method.
These methods allowed for comprehensive model training, validation, and rigorous evaluation of classification performance.

3. Results

3.1. Feature selection and model training

Feature Selection Wilcoxon rank-sum tests were applied to identify genes significantly differentially expressed between malignant and non-malignant samples. The p-values generated for all 5,000 genes were ranked, and the top 30 most significant genes were selected for downstream analysis. Notably, these included key markers such as HSPB1, SCD5, and MT2A.

Model Training and Evaluation The dataset, composed of expression data from the top 30 genes, was randomly split into a training set (60%) and a testing set (40%). Four classification models—Random Forest, Neural Network (MLP), Support Vector Machine (SVM), and XGBoost—were trained and evaluated.

3.2. Performance of the random forest classifier

In table1,the Random Forest classifier achieved the highest accuracy (89.93%) and excellent discriminative ability (ROC AUC = 0.947) [3]. The model demonstrated strong precision (87-92%) and recall (89-91%) across both classes, reflecting a balanced classification performance. Notably, the false-negative rate (FN=654) was lower than other models, indicating superior sensitivity in identifying malignant cells.

Table 1. The results of the random forest classifier

	precision	recall	f1-scorfe	support
0	0.92	0.89	0.91	9013
1	0.87	0.91	0.89	6998
Accuracy			0.9	16011
Macro avg	0.9	0.9	0.9	16011
Weighted avg	0.9	0.9	0.9	16011
Confusion matrix	[[8055 958]	[654 6344]]		

DOI: 10.54254/2753-8818/2025.LD27586

3.3. Performance of the MLP classifier

In Table2,MLP classifier achieved an overall accuracy of 87%, with precision ranging from 85% - 89% and recall spanning 86% - 88% across the two classes. The f1 - scores for the two classes were 0.86 and 0.89 respectively, and both macro average and weighted average of f1 - score reached 0.87, reflecting a relatively balanced classification performance. From the confusion matrix, the false - negative count (FN = 985) and false - positive count (FP = 1048) showed that there was still room for improvement in distinguishing between the two classes. Specifically, the precision of class 0 was 0.89, recall was 0.88, indicating that while it had a good ability to correctly identify class 0 samples, there was a certain misclassification rate; for class 1, precision was 0.85 and recall was 0.86, suggesting that the model's recognition of class 1 samples was slightly weaker in precision compared to class 0. Overall, the model showed a decent classification performance but could be further optimized to reduce misclassifications.

Table 2. The results of the random forest classifier

	precision	recall	f1-scorfe	support
0	0.89	0.88	0.89	9013
1	0.85	0.86	0.86	6998
Accuracy			0.87	16011
Macro avg	0.87	0.87	0.87	16011
Weighted avg	0.87	0.87	0.87	16011
Confusion matrix	[[7965 1048]	[985 6015]]		

3.4. Performance of the xgboost classifier

In table 3,the XGBoost classifier performed similarly well, yielding an accuracy of 89.8% and the highest ROC AUC score (0.950), highlighting excellent overall predictive capability. Precision and recall rates for this model were consistently high (approximately 88-91%), suggesting balanced effectiveness.

Table 3. The results of XGBoost classifier

	precision	recall	f1-scorfe	support
0	0.91	0.88	0.9	9013
1	0.85	0.89	0.87	6998
Accuracy			0.88	16011
Macro avg	0.88	0.88	0.88	16011
Weighted avg	0.89	0.89	0.88	16011
Confusion matrix	[[7946 1067]	[780 6218]]		

3.5. Performance of the SVM classifier

In table 4,the SVM classifier showed moderately strong performance with an accuracy of 88.5% and ROC AUC of 0.938, demonstrating balanced precision and recall rates. The Neural Network (MLP) had slightly lower performance, achieving an accuracy of 87.3% and ROC AUC of 0.929, with a

DOI: 10.54254/2753-8818/2025.LD27586

higher false-negative rate indicating potential risks in diagnostic settings.Visualizations, including ROC curves and confusion matrix heatmaps, reinforced these findings, underscoring the Random Forest and XGBoost models as particularly suitable for clinical applications requiring accurate malignant cell identification.

Table 4. The results of SVM classifier

	precision	recall	f1-scorfe	support
0	0.91	0.91	0.91	9013
1	0.88	0.88	0.88	6998
Accuracy			0.9	16011
Macro avg	0.9	0.9	0.9	16011
Weighted avg	0.9	0.9	0.9	16011
Confusion matrix	[[7946 1067]	[780 6218]]		

4. Discussion and conclusion

Our study demonstrated that differential gene screening combined with ensemble learning methods, particularly Random Forest, efficiently classified malignant and non-malignant samples (accuracy ~90%, ROC AUC ~0.95) [4]. Key differentially expressed genes, such as MT2A, provide valuable insights and may guide future research on cancer biomarkers. Among the evaluated models, the Random Forest classifier exhibited the best overall accuracy and sensitivity, making it highly suitable for clinical diagnostic purposes. XGBoost also showed strong performance, suggesting its potential applicability in scenarios requiring robust generalization. Although the Neural Network (MLP) model displayed lower performance, further optimization might enhance its effectiveness.Despite these promising results, our study has certain limitations. The selection of the top 30 genes may omit other biologically relevant markers due to the stringent selection criteria. Additionally, the inherent imbalance in sample distribution between malignant and non-malignant categories may influence model performance. Future studies should address these limitations through more comprehensive feature selection strategies and techniques such as synthetic minority over-sampling (SMOTE) to better handle data imbalance.Looking forward, integrating additional data types, such as genomic variants or epigenetic modifications, may further enhance the classification accuracy and clinical utility of machine learning models. Combining single-cell transcriptomics with spatial transcriptomics could also provide deeper insights into the tumor microenvironment and interactions within GBM. Such integrated approaches promise to significantly advance personalized diagnostics and therapeutic strategies in glioblastoma research [5].

References

[1] Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., ... & Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science, 344(6190), 1396–1401. https: //doi.org/10.1126/science.1254257.

[2] Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M. J., Melms, J. C., ... & Regev, A. (2018). A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell, 175(4), 984–997. https: //doi.org/10.1016/j.cell.2018.09.006

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https:

DOI: 10.54254/2753-8818/2025.LD27586

//doi.org/10.1145/2939672.2939785

[4] Neftel, C., Laffy, J., Filbin, M. G., Hara, T., Shore, M. E., Rahme, G. J., ... & Bernstein, B. E. (2019). An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell, 178(4), 835–849.e21. https: //doi.org/10.1016/j.cell.2019.06.024

[5] Rao, A., Barkley, D., França, G. S., & Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. Nature, 596(7871), 211–220. https: //doi.org/10.1038/s41586-021-03634-9