Proceedings of ICEGEE 2026 Symposium: AI-Based Medicine and Biological Data Analysis
DOI: 10.54254/2753-8818/2026.AU34017

Machine Learning-Based Classification of Malignant Cells in Glioblastoma Using Single-Cell RNA Sequencing Data

Yixiang Chen

School of Mathematics and Statistics, The University of Sydney, Sydney, Australia
15529210000@163.com

Abstract. Glioblastoma (GBM) is an aggressive primary brain tumour marked by pronounced cellular diversity. In single-cell RNA sequencing (scRNA-seq) studies, distinguishing malignant cells from surrounding non-malignant cells is essential for interpreting tumour biology and estimating tumour purity, but manual annotation is often slow and subjective. In this work, machine learning models were trained on a curated scRNA-seq dataset comprising 40,026 cells and 5,000 genes. Genes were screened with Welch's t-test, and the 30 most informative features were retained for model development. The data were split into training (60%), validation (20%), and independent test (20%) subsets. An Extreme Gradient Boosting (XGBoost) classifier was compared with a feedforward neural network. On the independent test set, XGBoost reached an area under the ROC curve (AUC) of 0.9529, exceeding the neural network result (AUC = 0.9136). It also maintained well-balanced sensitivity and specificity. Moreover, the proportion of cells predicted as malignant offered a useful proxy for tumour purity. Overall, the results indicate that gradient boosting is a practical and scalable option for automated malignant-cell identification in GBM single-cell data.

Keywords: glioblastoma, single-cell RNA sequencing, malignant cell classification, XGBoost, tumor purity

1. Introduction

Glioblastoma (GBM) is widely recognised as the most aggressive primary brain tumour in adults. Even with multimodal treatment, patient outcomes remain poor, largely because GBM shows substantial intratumoural heterogeneity and marked resistance to therapy [1, 2].

The emergence of single-cell RNA sequencing (scRNA-seq) has made it possible to examine tumour ecosystems at much finer resolution than bulk transcriptomic approaches [3]. In glioma research, single-cell studies have shown that malignant cells can coexist with diverse stromal and immune populations within the same specimen, which makes cell-level classification especially important [2, 4].

Reliable identification of malignant cells matters for several reasons: it improves biological interpretation, supports downstream analyses, and can inform estimates of tumour purity. However, manual labelling based on canonical markers is not always consistent across studies and can become

DOI: 10.54254/2753-8818/2026.AU34017

inefficient when the number of cells is large. For that reason, a data-driven classification framework is a useful addition to current analytical workflows.

In this study, machine learning models were developed and evaluated to classify malignant and non-malignant cells in GBM using curated scRNA-seq data from Couturier et al. [2]. The study further explored whether cell-level predictions could be summarised to provide a simple estimate of tumour purity, in line with earlier work emphasising the importance of purity assessment in cancer transcriptomic analysis [5].

2. Methodology

2.1. Dataset

The dataset analysed here came from the GBM single-cell RNA sequencing study reported by Couturier et al. [2]. The curated expression matrix included 5,000 genes measured across 40,026 individual cells. In this matrix, rows corresponded to genes and columns corresponded to cells. The expression data had already been normalised, and a value of zero indicated that no expression was detected for a given gene in a given cell.

Cells were annotated as malignant cells (n = 17,403), tumor-adjacent brain cells (n = 14,705), or brain cells from normal donors (n = 7,918). For classification, the latter two groups were combined into a single non-malignant category. Quality filtering and normalization had already been completed before analysis.

2.2. Feature selection

To make the models easier to interpret and to reduce the dimensionality of the input space, an initial univariate filtering step was applied before model fitting. For each gene, expression levels in malignant and non-malignant cells were compared with Welch's t-test, which is appropriate when equal group variances cannot be assumed. Genes were processed iteratively so that memory demands remained manageable during computation.

Multiple-testing correction was carried out with the Benjamini-Hochberg procedure, after which the 30 most significant genes were retained as features for downstream modelling.

2.3. Data splitting and modelling

The full dataset was randomly partitioned into training (60%), validation (20%), and independent test (20%) subsets. Class proportions were preserved during splitting. The training subset was used to estimate model parameters, the validation subset supported hyperparameter tuning and threshold selection, and the test subset was held back until the final evaluation stage to provide an unbiased performance assessment.

Two principal classifiers were considered: XGBoost and a feedforward neural network. XGBoost is a tree-boosting algorithm that has shown strong performance on structured prediction tasks [6]. For the XGBoost model, the final classification threshold was chosen on the validation set using Youden's J statistic so that sensitivity and specificity were balanced. A baseline logistic regression model was also included in the comparative probability plots to provide an additional point of reference.

DOI: 10.54254/2753-8818/2026.AU34017

2.4. Model evaluation

Performance on the independent test set was assessed with receiver operating characteristic (ROC) curves, the area under the ROC curve (AUC), sensitivity, specificity, and the distribution of predicted probabilities.

3. Results

3.1. Dataset overview

The curated dataset contained 40,026 cells and 5,000 genes. Of these, 17,403 were labelled as malignant and 22,623 as non-malignant. The predefined split preserved the original class balance across the training, validation, and test subsets. Figure 1 shows that non-malignant cells were slightly more common than malignant cells in the full dataset.

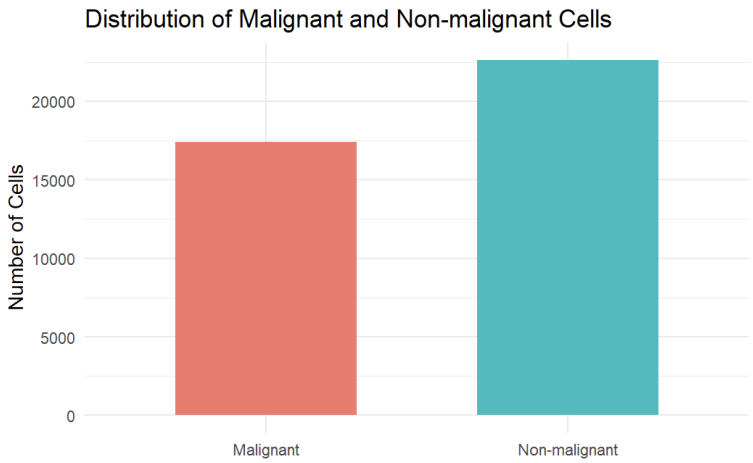

Figure 1. Distribution of malignant and non-malignant cells in the curated GBM single-cell dataset

3.2. Differential gene selection

Welch's t-test identified genes showing substantial expression differences between malignant and non-malignant cells. After Benjamini-Hochberg adjustment, the top 30 genes were kept as classification features. As illustrated in Figure 2, the volcano plot shows a clear differential-expression pattern, which supports the use of these genes in the modelling step.

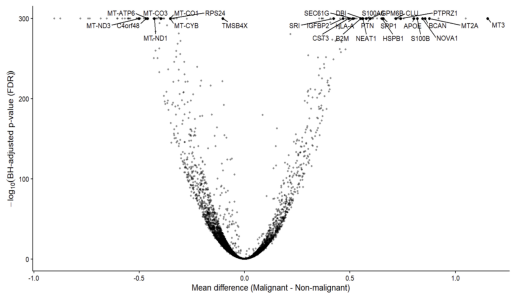

Figure 2. Volcano plot of differential expression between malignant and non-malignant cells based on Welch's t-test. Each point represents one gene, and the top 30 selected genes are highlighted and labelled

DOI: 10.54254/2753-8818/2026.AU34017

3.3. Model performance

On the independent test set, the XGBoost classifier displayed strong discriminative performance, with an AUC of 0.9529. When the validation-derived threshold was applied, the model achieved a sensitivity of 0.9066 and a specificity of 0.9019. These values indicate that the classifier performed well for both malignant and non-malignant cells. Figure 3 likewise shows an ROC curve concentrated near the upper-left corner, consistent with effective class separation.

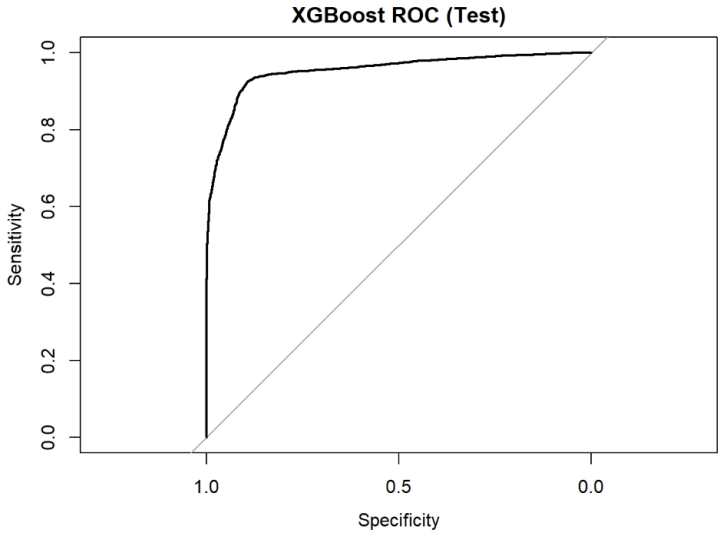

Figure 3. Receiver operating characteristic curve of the XGBoost classifier on the independent test dataset

The neural network also produced good test performance, although it did not match the XGBoost model. Its test AUC was 0.9136, which still indicates useful discrimination but with weaker separation than the boosting approach. As shown in Figure 4, the ROC curve remains clearly above the diagonal baseline, though the shape is less pronounced than that observed for XGBoost.

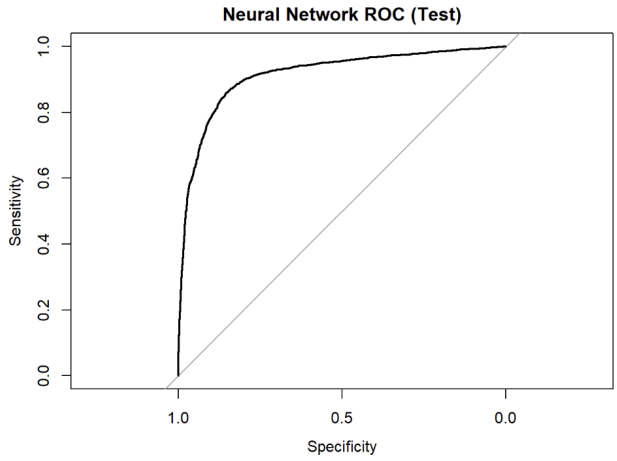

Figure 4. Receiver operating characteristic curve of the neural network classifier on the independent test dataset

DOI: 10.54254/2753-8818/2026.AU34017

A comparison of predicted malignant probabilities across models indicated that XGBoost generated the clearest distinction between the two cell classes. Most malignant cells received high predicted probabilities, whereas most non-malignant cells were assigned low values. Figure 5 shows that this separation was sharper for XGBoost than for logistic regression or the neural network.

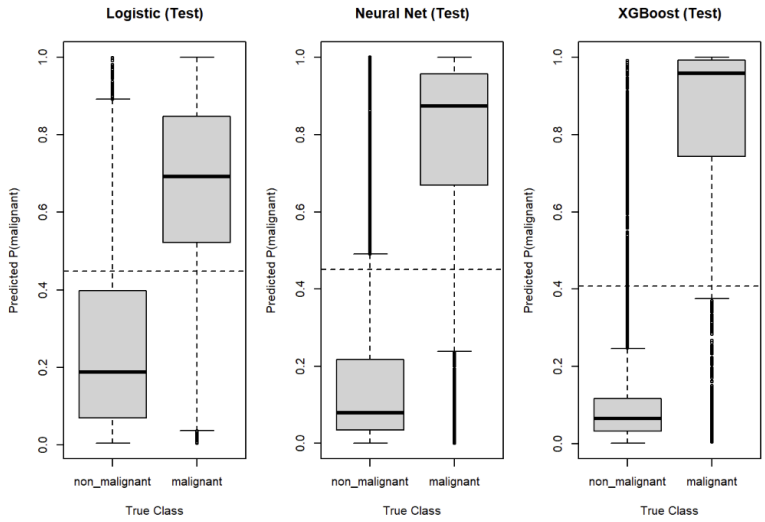

Figure 5. Comparison of predicted malignant probabilities across logistic regression, neural network, and XGBoost models on the independent test dataset

3.4. Tumor purity estimation

Tumour purity was estimated as the proportion of cells predicted to be malignant in the dataset. The resulting estimate was close to the annotated malignant-cell proportion, suggesting that the proposed classification framework may be useful for rapid, automated purity assessment.

4. Discussion

This study shows that malignant and non-malignant GBM cells can be separated with high accuracy using a relatively small panel of selected genes. Among the models evaluated, XGBoost performed better than the feedforward neural network, suggesting that the boosting framework captured complex, potentially nonlinear patterns in the expression data more effectively.

The value of the framework is not limited to binary cell classification. By aggregating cell-level predictions into a malignant-cell proportion, the model also provides a straightforward computational route for approximating tumour purity. That is relevant because impurity from surrounding non-malignant cells can complicate interpretation in downstream genomic analyses [5].

Several limitations should be noted. First, the analysis relied on a single curated dataset, so the generalisability of the findings should be checked on additional independent GBM cohorts. Second, feature selection was based on univariate testing, which may miss multivariate expression patterns. Third, although the neural network was reasonably competitive, this study did not systematically explore deeper architectures or more advanced regularisation strategies.

Future work could strengthen the proposed pipeline through cross-dataset validation, alternative feature-selection strategies, and the integration of complementary molecular signals such as copy number variation. Applying the same framework to other tumour types may also help extend its usefulness for broader tumour microenvironment studies.

DOI: 10.54254/2753-8818/2026.AU34017

5. Conclusion

Machine learning offers an effective and scalable way to identify malignant cells in GBM single-cell RNA sequencing data. In this study, XGBoost delivered the strongest predictive performance and produced a biologically plausible estimate of malignant-cell prevalence. Taken together, these findings support the use of gradient boosting as a practical method for automated malignant-cell classification and tumour purity assessment in curated GBM scRNA-seq datasets.

References

[1] Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science, 344(6190), 1396-1401.
[2] Couturier, C. P., Ayyadhury, S., Le, P. U., Nadaf, J., Monlong, J., Riva, G., et al. (2020). Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nature Communications, 11(1), 3406.
[3] Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
[4] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods, 6(5), 377-382.
[5] Neftel, C., Laffy, J., Filbin, M. G., Hara, T., Shore, M. E., Rahme, G. J., et al. (2019). An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell, 178(4), 835-849.e21.
[6] Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications, 4, 2612.