Glioma classification and risk analysis

1. Introduction

Gliomas represent a significant challenge in neuro-oncology, accounting for most primary brain tumors in adults (AANS, 2024). The World Health Organization (WHO) classification system for gliomas, updated in 2021, provides a detailed framework for tumor classification based on histopathological and molecular features (Louis *et al*., 2021). This classification system incorporates advanced genetic and epigenetic markers, providing more understanding of glioma subtypes and their prognostic implications.

In recent years, machine learning (ML) techniques have emerged as powerful tools for medical diagnostics, including the classification of gliomas, taking advantage of large datasets to identify patterns and predict tumor characteristics with high accuracy (Pereira et al., 2016). Traditional ML models have primarily focused on improving the accuracy of glioma classification based on the WHO 2021 framework. However, there remains a significant opportunity to enhance these models by integrating additional prognostic factors, particularly survival analysis.

Previous studies have demonstrated the efficacy of machine learning in classifying gliomas using various features such as imaging data, genetic profiles, and histopathological information (BSE, 2020; Zhang et al., 2021). However, these approaches often fall short in integrating survival data. The integration of survival risk analysis with classification models has been limited, with most existing models focusing primarily on classification accuracy rather than outcome prediction.

This project presents a novel approach by combining advanced machine learning techniques for glioma classification, according to the WHO 2021 criteria, with an analysis of survival risk. This dual approach addresses a critical gap in current methodologies, which typically do not provide a comprehensive risk assessment, offering a more comprehensive tool for clinical decision-making and personalized treatment planning.

2. Methodology

2.1 Data Collection

The datasets used with the clinical and DNA methylation information were from The Cancer Genome Atlas (TCGA). The dataset included clinical data from 11160 patients, and the gene data included 450k genes of the patients.

2.2 Data pre-processing

The clinical database includes patients with all types of cancer. Therefore, patients were filtered to include only those classified with glioblastoma (GB) or low-grade glioma (LG), the latter encompassing astrocytoma and oligodendroglioma, as the database is categorized according to the WHO 2016 classification. Subsequently, the clinical database was merged with the genomic data using the patient ID, removing columns with more than 50 missing values, as well as rows, resulting in a dataset with approximately 403,973 genes out of the original 450,000.

Finally, a mapping was performed to identify glioma subtypes, as follows:

- Astrocytoma: Class 0
- Glioblastoma: Class 1
- Oligodendroglioma: Class 2

2.3 Classifier

Two models were used: Logistic Regression and Random Forest. The dataset was divided into testing (20%) and training (80%) subsets.

2.3.2 Explainable artificial intelligence (XAI)

The SHAP library was used to calculate SHAP values, which indicate which features have the most impact on the classifier models.

2.4 Risk analysis

For the risk analysis, three main events included in the clinical database were considered: overall survival event (OS), disease-specific survival event (DSS), and progression-free interval event (PFI), as well as the time of each. The original database also included disease-free interval event (DFI); however, it was decided not to use it due to the high number of missing values. The Logistic Regression and XGB Classifier models were used.

3. Results

3.1 Classifier

The following accuracy values were obtained for the proposed models:

Table 1. **Results of the classifier models**		
Accuracy	**Logistic Regression**	**Random Forest**
Overall	0.99	0.98
Class 0	1.00	0.98
Class 1	1.00	1.00

Class 2	0.97	0.97

3.2 XAI

After calculating the SHAP values, the most relevant features for the models were identified, resulting in a table for comparison across the three subtypes.

Table 2. **XAI for Logistic Regression**		
Astrocytoma	**Glioblastoma**	**Oligodendroglioma**
cg24597705	cg25594899	cg18115132
cg03909781	cg06097659	cg15814736
cg25594899	cg27561954	cg05376227
cg27561954	cg03909781	cg24189559
cg16991768	cg23552821	cg21943117
cg06097659	cg16991768	cg18572219
cg21813376	cg16619049	cg26155520
cg21915799	cg15988843	cg23072823
cg11407801	cg10375192	cg00369438
cg23519022	cg22606681	cg11404544
cg22606681	cg24597705	cg19823490
cg17393296	cg13461447	cg12633102
cg26798702	cg08750554	cg15185794
cg15728256	cg02587648	cg10049708
cg15988843	cg26314055	cg25610492
cg23552821	cg27181471	cg12461469
cg04218345	cg03258665	cg05526341
cg02587648	cg21288685	cg19865916
cg17441401	cg01412518	cg24553170
cg08748615	cg21913897	cg08748615

Table 3. **XAI for Random Forest**		
Astrocytoma	**Glioblastoma**	**Oligodendroglioma**
cg08765301	cg16328207	cg08765301
cg17737263	cg14355794	cg17737263
cg12978275	cg08836619	cg12978275
cg06967124	cg26812852	cg06967124
cg16328207	cg26334299	cg06561366
cg21545988	cg17961327	cg21545988
cg06561366	cg22114991	cg23643330
cg08836619	cg01906848	cg24880387

cg02580900	cg23368159	cg18079128
cg21913974	cg24360174	cg05650171
cg24360174	cg05195756	cg13101087
cg23643330	cg14718848	cg07516252
cg00885461	cg15224059	cg01565608
cg00529567	cg02210967	cg15126733
cg00857851	cg02876211	cg24397241
cg02210967	cg24731625	cg00350405
cg01906848	cg00885461	cg13856825
cg05987787	cg11643186	cg04567600
cg26377276	cg04682802	cg18735519
cg17667595	cg04059647	cg08362738

3.3 Risk analysis

Table 4. **Risk analysis using overall survival event (OS)**		
Score	**Logistic Regression**	**XGBClassifier**
Overall Model on Training Set: Overall Model on Test Set:	0.989 0.991	0.989 0.991
Year 1 Model on Training Set: Model on Test Set:	 0.944 0.905	 0.957 0.905
Year 3 Model on Training Set: Model on Test Set:	 0.959 0.940	 0.961 0.931
Year 5 Model on Training Set: Model on Test Set:	 0.961 0.966	 0.972 0.948
Year 7 Model on Training Set: Model on Test Set:	 0.968 0.966	 0.968 0.966

Table 5. **Risk analysis using disease-specific survival event (DSS)**		
Score	**Logistic Regression**	**XGBClassifier**
Overall Model on Training Set: Overall Model on Test Set:	0.998 1.000	0.987 1.000
Year 1 Model on Training Set: Model on Test Set:	 0.950 0.871	 0.968 0.871
Year 3		

Model on Training Set: Model on Test Set:	0.952 0.922	0.955 0.914
Year 5 Model on Training Set: Model on Test Set:	 0.972 0.957	 0.957 0.966
Year 7 Model on Training Set: Model on Test Set:	 0.976 0.974	 0.970 0.957

Table 6. **Risk analysis using progression-free interval event (PFI)**		
Score	**Logistic Regression**	**XGBClassifier**
Overall Model on Training Set: Overall Model on Test Set:	0.846 0.863	0.846 0.836
Year 1 Model on Training Set: Model on Test Set:	 0.842 0.836	 0.922 0.853
Year 3 Model on Training Set: Model on Test Set:	 0.816 0.853	 0.816
Year 5 Model on Training Set: Model on Test Set:	 0.846 0.793	 0.911 0.784
Year 7 Model on Training Set: Model on Test Set:	 0.835 0.836	 0.835 0.836

4. Discussion

The results of this study indicate that both Logistic Regression and Random Forest models show strong performance in glioma classification, with accuracies approaching 99% overall. This suggests that these models are highly effective in differentiating between glioma subtypes such as astrocytoma, glioblastoma, and oligodendroglioma. Specifically, Random Forest and Logistic Regression achieved nearly identical accuracies, highlighting the robustness of both methods for classification tasks.

In terms of feature importance, the SHAP analysis results for both the Logistic Regression and Random Forest models highlight the most influential genes for each glioma subtype. It was revealed that certain DNA methylation markers are more significant for classification, with differences in feature importance across subtypes. This can help in understanding which

molecular features are key for distinguishing between glioma types, providing valuable information for both diagnosis and research. Notably, there are no shared genes between Oligodendroglioma and Glioblastoma in either model. The absence of shared genes could reflect the fundamentally different biological pathways or molecular mechanisms. These findings could indicate that each glioma subtype has a unique genetic signature and might respond differently to treatments or have different disease progressions. The discrepancy in gene importance between the two models may be attributed to how each algorithm interprets and processes the data.

For the risk analysis, which assesses survival outcomes, the models performed well but showed varying results depending on the type of survival event. The Logistic Regression and XGB Classifier models demonstrated high accuracy for overall survival (OS) and disease-specific survival (DSS) events, particularly for later time points. However, performance was less consistent for progression-free interval (PFI) events, with slightly lower accuracy rates. This discrepancy highlights the complexity of predicting PFI compared to overall and disease-specific survival, likely due to the varying nature of progression in glioma patients or due to the number of missing values in the dataset.

5. Conclusions

In summary, the study successfully demonstrates that advanced machine learning models, specifically Logistic Regression and Random Forest, can accurately classify glioma subtypes and integrate survival risk analysis. Both models showed high classification accuracy, indicating their potential for improving diagnostic accuracy in clinical settings. The SHAP values provided a deeper understanding of feature importance, which can guide future research and clinical practices.

The risk analysis component of the study indicates that while the models are strong predictors for overall and disease-specific survival, there is room for improvement in predicting progression-free intervals. Future work could focus on refining these models or exploring additional features to enhance the prediction of progression-free survival. Overall, integrating machine learning with survival risk analysis represents a significant advancement in glioma prognosis and could lead to more personalized treatment strategies for patients.

6. References

- AANS. (2024). *Brain Tumors*. American Association of Neurological Surgeons. https://www.aans.org/patients/conditions-treatments/brain-tumors/
- BSE. (2020). *Machine Learning for Glioma Classification: A Review*. Journal of Neuro-Oncology, 149(2), 237-247.
- Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng HK, Pfister SM, Reifenberger, G., Soffietti, R., von Deimling, A., Ellison, D.W. (2021). *The 2021 WHO Classification of Tumours of the Central Nervous System: A summary*. Neuro-Oncology, 23(8):1231-1251. doi: 10.1093/neuonc/noab106. PMID: 34185076; PMCID: PMC8328013.
- Zhang, Y., Yang, J., & Lu, H. (2021). *Deep Learning Approaches for Glioma Classification*. Medical Image Analysis, 68, 101-113.