Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(1), 20-32
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Comparative Analysis of ML Models for Survival Prediction of Glioblastoma

Muna Awel[1,2], Dave Rushit[2], Samantha J. Katner[1], Mansi Bhavsar[2]

[1]Department of Biochemistry, Chemistry, and Geology, Minnesota State University, Mankato, Mankato, MN, USA
[2]Department of Computer Information Science, Minnesota State University, Mankato, Mankato, MN, USA
Email: muna.awel@mnsu.edu, rushit.dave@mnsu.edu, samantha.katner@mnsu.edu, mansi.bhavsar@mnsu.edu

How to cite this paper: Awel, M., Rushit, D., Katner, S.J. and Bhavsar, M. (2026) Comparative Analysis of ML Models for Survival Prediction of Glioblastoma. *Journal of Computer and Communications*, **14**, 20-32.
https://doi.org/10.4236/jcc.2026.141002

Received: December 16, 2025
Accepted: January 16, 2026
Published: January 19, 2026

Abstract

Glioblastoma multiforme (GBM) remains one of the most aggressive brain malignancies, with a median survival of less than 15 months. This study advances glioblastoma multiforme (GBM) survival prediction by developing a comprehensive machine learning (ML) pipeline that integrates four classifiers: Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM) on TCGA-derived multi-omics datasets. Rigorous preprocessing, including missing data assessment (MCAR test), multicollinearity checks, and feature selection, was followed by hyperparameter optimization using GridSearchCV and 10-fold cross-validation to enhance model performance and generalizability. Predictive performance was evaluated with AUC-ROC, precision-recall curves, and classification reports, while interpretability was assessed through SHAP (SHapley Additive exPlanations) analysis to identify the most influential features driving survival predictions. Random Forest achieved the highest predictive accuracy while maintaining strong interpretability, highlighting key drivers of GBM prognosis such as age, MGMT promoter methylation status, and specific gene expression signatures. Despite promising results that demonstrate ML's ability to handle GBM heterogeneity, limitations include the relatively modest sample size and lack of external validation. Future work will incorporate independent cohorts for external validation, explore advanced ensemble and hybrid modeling strategies, and further optimize models to meet clinical requirements for both accuracy and transparent decision-making.

Keywords

Glioblastoma Multiforme, Machine Learning, Survival Prediction, Multi-Omics, Methylation Biomarkers, GridSearchCV, SHAP, Interpretability, ROC Curve

1. Introduction/Background Study

Glioblastoma multiforme (GBM) is well known as the most aggressive and prevalent primary malignant brain tumor. GBM accounts for approximately 48.6% of all malignant central nervous system tumors [1] [2]. Recently, glioblastoma was classified as a high-grade glioma (grade IV) based on the molecular mutation profiles such as IDH mutant/wildtype, and CDKN2A/B homozygous deletion, as well as whether necrosis or microvascular proliferation is observed [3]. Lower-grade diffuse gliomas do not demonstrate the same degree of biologic aggressiveness, rapid progression, or resistance to current therapies as GBM [4]. Thus, among all malignant primary brain tumors, GBM remains the most common and lethal subtype and contributes to more than 15,000 deaths annually in the United States [4]. The median overall survival remains dismal at 14 - 15 months after surgical diagnosis, even with standard temozolomide-radiotherapy regimens [5].

Despite aggressive multimodal therapy, long-term outcomes for patients with GBM remain extremely poor, highlighting the limitations of current clinical decision-making guided primarily by histopathological classification and treatment response [3] [4]. Moreover, GBM exhibits marked inter- and intra-tumor heterogeneity driven by complex genetic and epigenetic alterations that contribute to variable disease evolution and therapeutic resistance [6]. Recent large-scale network-based analyses of tumor genetics have shown that high-dimensional graph representations can more accurately capture survival signatures than traditional diagnostic categories alone [6]. Complementary work integrating single-cell RNA sequencing, spatial transcriptomics, and deep learning on whole-slide histology images has further demonstrated that spatial cellular architecture and transcriptional subtype composition are strongly associated with prognosis in GBM, and that these features can be inferred directly from routine histology [7]. Together, these epidemiologic and molecular insights underscore the urgent need for robust prognostic biomarkers and advanced predictive modeling approaches capable of supporting personalized management in GBM.

One of the most clinically significant sources of heterogeneity is DNA methylation, which strongly influences tumor progression, therapeutic response, and prognosis [8]. For instance, mutations in the isocitrate dehydrogenase (IDH)1/2 genes promote accumulation of the oncometabolite 2-hydroxyglutarate, inducing genome-wide hypermethylation and improved clinical outcomes in IDH-mutant GBM [9]. Another subtype, the methylation of the O6-methylguanine-DNA methyltransferase (MGMT) gene promoter, silences this DNA repair enzyme [9]. Therefore, methylated MGMT subtype GBM tumors are more to alkylating agents and serve as another favorable prognostic biomarker associated with extended survival. Additionally, the methylated MGMT status has a predictive value with higher cutoffs in IDH-mutant tumors [10]. TMZ, a standard-of-care oral alkylating agent administered concurrently with radiotherapy and followed by maintenance cycles, significantly improves median survival compared to radiotherapy alone, particularly in MGMT-methylated patients where it enhances treatment ef-

ficacy by impairing DNA repair mechanisms [10]. Thus, understanding the effects of different therapies is important to find the best mechanism that is beneficial to the patient.

With advancement of technologies in machine learning (ML) and deep learning newer mechanisms have developed integrating multi-omics data, radiomics, and clinical variables. Ensemble methods, such as gradient-boosted trees and deep neural networks, have improved predictive performance by addressing data heterogeneity and enhancing interpretability via techniques like SHAP, tackling limitations in traditional models [11]. However, many existing GBM survival prediction studies primarily emphasize predictive accuracy, often relying on complex model architectures, while providing limited attention to interpretability, reproducibility, and consistent benchmarking across models. In addition, prior studies frequently evaluate models using heterogeneous datasets or preprocessing strategies, making direct comparison and clinical translation challenging. This research builds on these advancements by developing a comprehensive ML pipeline using logistic regression, random forest, SVM, and XGBoost on TCGA-derived multi-omics datasets, incorporating SMOTE for class imbalance and GridSearchCV for optimization, to predict binary survival outcomes.

2. Methodology

This section explores the comprehensive methodology employed within the study, covering topics on data acquisition, preprocessing, feature engineering, and model development. All analyses were conducted using Python (version 3.12) with libraries including scikit-learn (for logistic regression and imputation), XGBoost (for gradient boosting), SHAP (model interpretation). The methodology adheres to best practices in biomedical data science, prioritizing reproducibility, statistical rigor, and ethical considerations in handling sensitive health data.

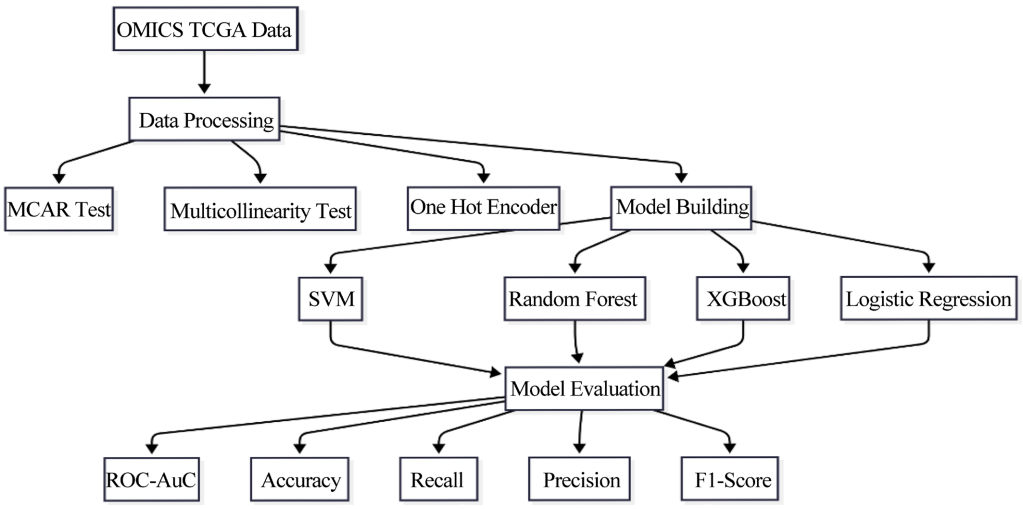

Figure 1. System architecture of proposed methodology highlights the sequential flow from data input to final model assessment, ensuring clarity in how each stage contributes to the overall system.

To visualize the workflow, a system architecture diagram (Figure 1) illustrates the sequential steps of the data pipeline, from raw data ingestion to model interpretation. This architecture ensures a modular, scalable design that can be adapted for similar omics-based studies.

2.1. Data Collection

The datasets utilized within this study are sourced from The Cancer Genome Atlas (TCGA) via cBioPortal, specifically the Glioblastoma Multiforme (TCGA, Cell 2013 and Firehose Legacy) datasets, comprising 577 samples with multi-omics data. EDA was performed to characterize feature distributions, identify data types, and detect anomalies.

2.2. Data Preprocessing

The data preprocessing step is the critical step in data analysis to ensure the data quality, integrity and prepare the data inputs for modeling. In fact, this involved several sub-steps executed in a pipeline using scikit-learn's Pipeline class for reproducibility.

2.3. Missing Values

Missing values were first quantified to identify the correct mechanism to handle the null values. Features that had missing values 50% or more were dropped to reduce noise and bias for the models. To characterize the missing data mechanism, a global Little's MCAR test was conducted rejecting the null hypothesis of complete randomness that missingness occurred completely at random [12]. To further explore dependencies, pairwise independence tests were performed between missingness indicators for all unique feature pairs. For categorical variables χ^2 tests of independence between the categorical value and the missingness indicators of other features. On the other hand, for continuous variables Welch's two-sample t-tests were performed comparing means across the missing/not-missing groups of other features as shown in Table 1 [13].

Table 1. t-test results for continuous variables vs. missingness indicators show significant dependencies ($P < 0.05$). The t-test results assess whether continuous clinical variables differ based on missingness in key molecular markers. The significant P-values indicate that missingness is not completely random and may reflect underlying patient or biological characteristics.

Continuous Variable	Missingness Variable	t-statistics	P-value
Diagnosis Age	IDH1 Mutation	−2.0759	0.0394
Overall Survival (Months)	Methylation Status	2.1364	0.0333
Overall Survival (Months)	MGMT Status	2.1831	0.0296

Table 1 shows the results of two sample t-tests comparing the means of continuous variables against the missingness of other variables, where significant differ-

ences ($P < 0.05$) were observed. The t-statistic indicates the magnitude and direction of the mean difference, with negative values suggesting a higher mean in the "missing" group, and positive values indicating a higher mean in the "not missing" group.

Overall based on the analysis of handling missing values the non-MCAR, likely MAR mechanism and clinical relevance, imputation was preferred. K-Nearest Neighbors (KNN) imputation was chosen to leverage multivariate relationships, suitable for mixed omics data [14]. This mechanism was selected over model-based approaches such as Multiple Imputation by Chained Equations (MICE) due to its non-parametric nature and its ability to preserve local multivariate structure without imposing distributional assumptions. This property is particularly advantageous for high-dimensional multi-omics datasets with mixed feature types, where nonlinear relationships and complex dependencies are common. Prior studies have demonstrated that KNN imputation performs competitively for genomic and epigenomic data under Missing at Random (MAR) mechanisms while maintaining computational efficiency and stability [15].

2.4. Handling Categorical Features

To enable numerical analysis and avoid introducing ordinal assumptions, one-hot encoding was applied to multi-class categorical variables using scikit-learn's OneHotEncoder. This process transformed each category into a binary column, creating a sparse matrix of dummy variables [16].

2.5. Multicollinearity Assessment

Variation Inflation Factor (VIF) was calculated for all features. VIF values of greater than 10 were dropped to mitigate high multicollinearity, which can destabilize model coefficients [17]. L2 regularization method (Ridge regression) was applied, penalizing large coefficients to stabilize model estimates [18]. This approach was implemented on the numeric dataset, including one-hot encoded variables, reducing the impact of correlated features.

2.6. Model Development

Prior to model training, Overall Survival (Months) originally recorded as a continuous variable was binarized using the cohort median survival time as the threshold, with patients surviving longer than or equal to the median labeled as long-term survivors (Class 1) and those below the median labeled as short-term survivors (Class 0). This binarization supports reproducible classification modeling and clinically meaningful risk stratification.

Logistic regression was the benchmark model for its high interpretability. Additional models like SVM, RF and XG BOOST are implemented for comparison to find a most suitable model. Regularization (L2) is applied to reduce multicollinearity and enhance model performance, while features are normalized using StandardScaler to ensure compatibility with scale-sensitive models [18]. Hyperpa-

rameter tuning is conducted using GridSearchCV to optimize model parameters, and 10-fold cross-validation with stratified sampling is employed to handle censored data and improve robustness [19]. This multi-model strategy, combined with preprocessing and tuning, aims to address gaps in interpretability and generalizability from prior studies.

2.7. Model Evaluation

This section evaluates model performance for binary GBM survival prediction. To address class imbalance, SMOTE was applied within training folds only to balance minority and majority outcomes. Performance is summarized by the ROC-AUC, reflecting discrimination across thresholds, and a classification report (accuracy, precision, recall, and F1) reported both per class and as macro-averages to capture minority-class behavior. Where relevant, PR-AUC complements ROC-AUC under imbalance. Estimates are obtained via stratified 10-fold cross-validation, preserving class proportions and yielding optimized, generalizable metrics.

3. Results

This section explores the results of each model and model interpretability of the features that were highly significant to the model's predictions.

Figure 2. A comparison of model performance across multiple evaluation metrics. This model presents CV ROC-AUC, Test ROC-AUC, Accuracy, Precision, Recall, and F1 scores for four classification models: Random Forest, XGBoost, Logistic Regression, and SVM.

Figure 2 compares the performance of four machine learning models Random Forest, XGBoost, Logistic Regression, and SVM across key evaluation metrics, including cross-validated ROC-AUC, test ROC-AUC, accuracy, and the average precision, recall, and F1-scores. Overall, Random Forest achieves the strongest performance, with the highest accuracy, precision, and F1-score, as well as competitive ROC-AUC values. XGBoost shows slightly lower accuracy than Random Forest but maintains comparable ROC-AUC and balanced average precision, re-

call, and F1. Overall, the chart highlights the trade-offs between models, showing how each one emphasizes different aspects of classification performance depending on whether accuracy, balance, or minority class detection is prioritized.

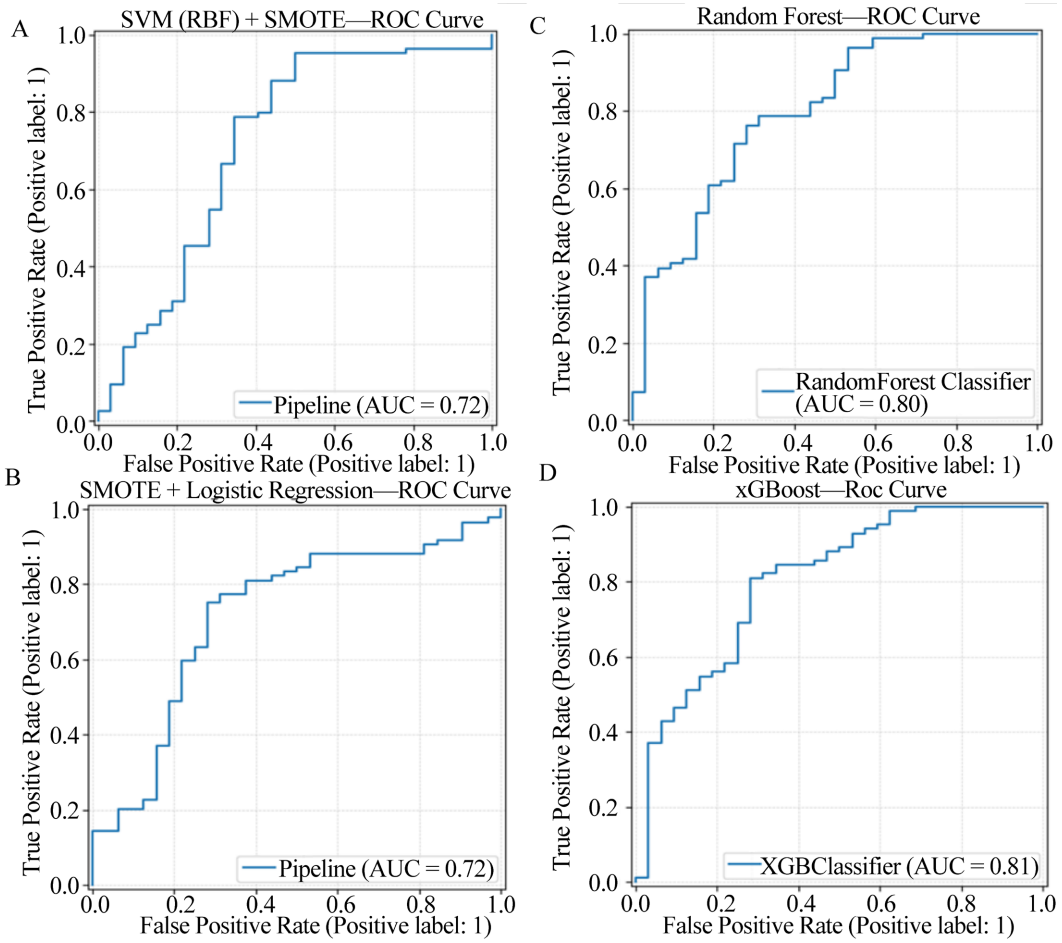

Figure 3. ROC curves for GBM survival classifiers. (A-B) ROC curves for Logistic Regression and SVM models trained with SMOTE to address class imbalance. (C-D) ROC curves for Random Forest and XGBoost models, which manage imbalance internally through their ensemble structures.

As shown in **Figure 3**, XGBoost achieved the highest AUC (0.81), closely followed by Random Forest (0.80), while Logistic Regression + SMOTE and SVM (RBF) + SMOTE both reached 0.72. Curves lie well above the diagonal, indicating discrimination better than chance.

4. Discussion

This study evaluated machine-learning approaches for binary GBM survival prediction using TCGA data. After rigorous preprocessing XGBoost achieved the highest ROC-AUC (~0.81), closely followed by Random Forest (~0.80), while SVM (RBF) and Logistic Regression reached ~0.72. These results suggest that models capturing nonlinear interactions and higher-order effects better reflect the heterogeneity of GBM.

Sensitivity (Recall)-Specificity Trade-off; in the aspect balancing between recall and specificity, minimizing false negative is important. Recall is true positive rate TP/(TP + FN) and increases as the threshold lowers, while specifically true negative rate for low-risk patients TN/(TN + FP) typically decrease. Thus, Random Forest supports a sensitivity-first strategy (very high Class-1 recall with acceptable precision) at the cost of more false positives (lower Class-0 recall/specificity), whereas XGBoost offers a more conservative balance (higher specificity at the same target sensitivity, with modestly lower recall).

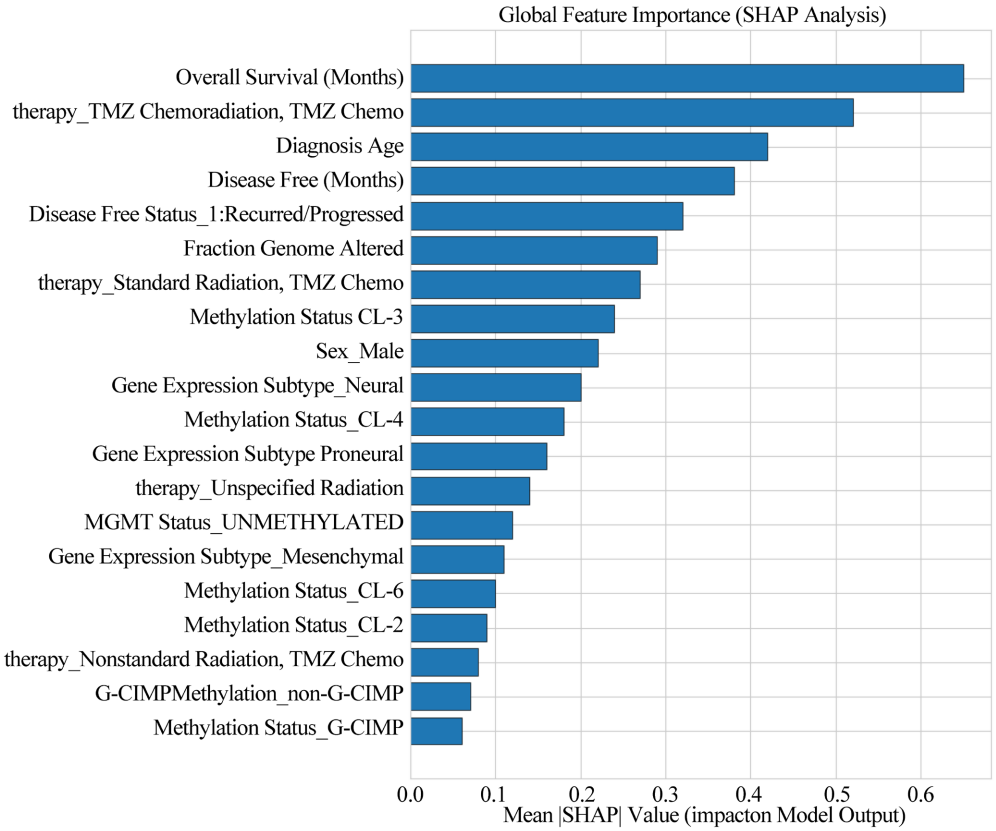

Figure 4. Global feature importance based on SHAP values. This bar chart summarizes the global contribution of each feature to the model's predictions using the mean absolute SHAP value (mean |SHAP|) across all samples. Features are ranked from top to bottom according to their overall impact on the model, with larger values indicating greater influence on prediction outcomes. It provides an aggregated view of feature importance across the cohort. Additionally, this representation facilitates straightforward comparison of feature importance and highlights the dominant clinical and molecular drivers underlying the model's prognostic decisions.

Model interpretation/Genomic Analysis: SHAP indicated that clinical/treatment variables and genome-level summaries contribute meaningfully to discrimination. Survival-proximal covariates overall survival months and disease-free duration, and therapy indicators exert the largest effects, with age and fraction genome altered (FGA) also contributing meaningfully. As shown in **Figure 4**, the therapy indicators show the treatment course for GBM patients. The feature "ther-

apy_TMZ Chemoradiation, TMZ Chemo" flags patients who received the modern standard Stupp regimen: concurrent external-beam radiation with **temozolomide (TMZ)** followed by adjuvant TMZ [20]. The "therapy_Standard Radiation, TMZ" denotes conventional-fractionation radiation co-administered with TMZ but recorded under a separate label; functionally this still reflects radiotherapy + TMZ [21]. The "therapy Nonstandard Radiation, TMZ Chemo" captures atypical radiotherapy schedules given with TMZ often used in older or frailer patients [22]. In the SHAP bar plot, therapy-related indicators exhibit high mean absolute SHAP values, indicating a strong global contribution to the model's predictive performance. Their prominence aligns with the known prognostic relevance of TMZ-based chemoradiation in many GBM cases [20]. In terms of DNA methylation, although MGMT promoter methylation and G-CIMP are well-established prognostic biomarkers, they show lower standalone SHAP importance in our model. This does not contradict their clinical relevance. Instead, it reflects on their redundancy with global methylation clusters (e.g., CL_3/CL_4/CL_6) that encode genome-wide CpG patterns and therefore absorb much of the MGMT/G-CIMP signal.

Clinical relevance: Although purely predictive, the models highlight variables aligned with current practice (e.g., age, therapy patterns) and may inform risk stratification for follow-up intensity or trial eligibility. Tree-based SHAP explanations can support clinician review by revealing case-level drivers of risk. For example, at the post-surgical evaluation stage, a patient receiving standard-of-care therapy but predicted by the model to be high risk based on combined molecular and clinical features could be considered for closer surveillance, earlier follow-up imaging, or prioritization for clinical trial enrollment. Tree-based SHAP explanations further support clinician review by revealing patient-specific drivers of risk, enabling predictions to be interpreted in the context of known biological and treatment-related factors rather than used as opaque risk scores.

5. Limitations

This study on glioblastoma multiforme (GBM) survival prediction using machine learning models presents several limitations that influence both the generalizability and clinical applicability of the findings. First, the dataset is relatively small, consisting of 577 multi-omics samples from TCGA. Although this cohort is well-curated, the limited sample size reduces statistical power and may not capture the full biological and clinical heterogeneity of GBM. Moreover, the study lacks external validation such as evaluation on CGGA or other independent datasets which restrict confidence in the model's generalizability across populations and sequencing platforms.

Second, certain preprocessing steps may have unintentionally weakened biologically relevant signals. For example, methylation preprocessing likely absorbed much of the G-CIMP and IDH information, leading these features to appear less significant in the SHAP interpretability analysis. This suggests that feature scaling

and transformation should be handled cautiously to preserve biologically meaningful variation. Additionally, the study relied on binary survival outcomes, which oversimplified time-to-event data and failed to account for censoring.

Third, while SMOTE was used to address class imbalance, synthetic oversampling may not accurately represent the intricate molecular and biological relationships of key biomarkers like IDH and MGMT. The resulting synthetic samples may distort data distribution and introduce artifacts, explaining the model's relatively weaker performance on the minority class. Future work should explore more advanced imbalance techniques, such as class-weighted learning or focal loss, to better handle uneven survival classes.

Interpretability also poses an important limitation. Although SHAP values were employed to improve transparency, models such as XGBoost and SVM remain largely "black-box", and SHAP explanations may not fully bridge the gap between algorithmic predictions and clinical trust. Clinical decision-making requires transparent and stable explanations, and further work integrating causally grounded or rule-based interpretability frameworks could strengthen reliability and physician confidence.

Another limitation lies in the scope of data modalities. Relying solely on multi-omics features overlooks other crucial determinants of GBM survival, such as radiographic imaging, surgical resection extent, treatment regimens, and patient performance status. Incorporating multimodal data including radiomics and clinical variables could yield a more comprehensive understanding of GBM heterogeneity and enhance model robustness. Overall, these limitations underscore the need for larger, externally validated, and multimodal datasets, more biologically informed preprocessing, survival-specific modeling strategies, and enhanced interpretability methods. Addressing these aspects in future research will be vital to improve transparency, and clinical adoption of machine-learning-based survival prediction models for GBM.

6. Conclusions

The central takeaway of this work is that predictive accuracy alone is insufficient for clinical relevance. By integrating SHAP-based explanations, the proposed framework links model predictions to biologically and clinically established factors such as IDH mutation and MGMT methylation, enabling model outputs to be reviewed and contextualized by clinicians rather than treated as opaque risk scores. This alignment between predictive modeling and domain knowledge is critical for building trust and supporting clinician-in-the-loop decision-making.

From a broader perspective, this work highlights how carefully benchmarked, interpretable ML pipelines can bridge the gap between computational performance and practical utility in neuro-oncology. Rather than advocating for fully automated decision-making, the framework illustrates a pathway for ML models to function as auxiliary decision-support tools, informing follow-up intensity, patient stratification, and research trial eligibility alongside standard clinical assessment.

Overall, these findings underscore the importance of balancing predictive power with interpretability in GBM survival modeling. By prioritizing transparency, reproducibility, and clinical alignment, this study contributes to the development of machine learning approaches that are not only technically robust but also positioned for meaningful integration into future precision oncology workflows.

7. Future Work

Future work will focus on strengthening the generalizability, interpretability, and clinical readiness of the proposed machine learning framework for GBM survival prediction. One of the most immediate steps is to incorporate external validation datasets and other multi-center cohorts, to evaluate model performance across diverse populations and sequencing platforms. This will help assess the reproducibility of predictive performance and mitigate potential dataset biases introduced by single-source training data. External validation will also enable testing of the pipeline under different demographic, genetic, and clinical distributions, ensuring that the model maintains its predictive reliability beyond the TCGA cohort.

Future research will also explore deep learning architectures to better model the complex, nonlinear relationships within high-dimensional omics data. Incorporating Convolutional Neural Networks (CNNs) and autoencoders can enable automatic feature extraction from multi-omics inputs, potentially improving representation learning and model scalability. Additionally, hybrid frameworks combining traditional ML algorithms with deep survival models will allow the incorporation of censored time-to-event data, providing more clinically relevant survival probabilities instead of binary outcomes.

Expanding the dataset beyond methylation and expression profiles is another important step. Integrating multi-modal data sources including radiomics, histopathological imaging, treatment histories, and demographic or clinical factors can improve the model's ability to capture the full heterogeneity of GBM. Such multimodal fusion can be achieved using late or intermediate fusion strategies, where features from different data types are merged to produce comprehensive survival predictions. This expansion will not only enhance model accuracy but also strengthen the biological interpretability of predictions.

Finally, emphasis will be placed on improving model deployment and real-world usability. Future iterations will include model calibration and uncertainty estimation through bootstrap confidence intervals and conformal prediction intervals to quantify prediction reliability. Building a lightweight deployment pipeline such as a web-based clinical interface or containerized application can enable real-time prediction and visualization within clinical workflows, particularly in resource-constrained environments where computational and data access limitations are prevalent. Ensuring reproducibility through open-source code, environmental documentation, and model cards will further promote transparency and facilitate collaboration among researchers and clinicians.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Qi, D., Li, J., Quarles, C.C., Fonkem, E. and Wu, E. (2022) Assessment and Prediction of Glioblastoma Therapy Response: Challenges and Opportunities. *Brain*, **146**, 1281-1298. https://doi.org/10.1093/brain/awac450

[2] Ostrom, Q.T., Price, M., Neff, C., Cioffi, G., Waite, K.A., Kruchko, C., *et al.* (2022) CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2015-2019. *Neuro-Oncology*, **24**, v1-v95. https://doi.org/10.1093/neuonc/noac202

[3] Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., *et al.* (2021) The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary. *Neuro-Oncology*, **23**, 1231-1251. https://doi.org/10.1093/neuonc/noab106

[4] Schaff, L.R. and Mellinghoff, I.K. (2023) Glioblastoma and Other Primary Brain Malignancies in Adults. *JAMA*, **329**, 574-587. https://doi.org/10.1001/jama.2023.0023

[5] Grochans, S., Cybulska, A.M., Simińska, D., Korbecki, J., Kojder, K., Chlubek, D., *et al.* (2022) Epidemiology of Glioblastoma Multiforme-Literature Review. *Cancers*, **14**, Article 2412. https://doi.org/10.3390/cancers14102412

[6] Ruffle, J.K., Mohinta, S., Pombo, G., Gray, R., Kopanitsa, V., Lee, F., *et al.* (2023) Brain Tumour Genetic Network Signatures of Survival. *Brain*, **146**, 4736-4754. https://doi.org/10.1093/brain/awad199

[7] Zheng, Y., Carrillo-Perez, F., Pizurica, M., Heiland, D.H. and Gevaert, O. (2023) Spatial Cellular Architecture Predicts Prognosis in Glioblastoma. *Nature Communications*, **14**, Article No. 4122. https://doi.org/10.1038/s41467-023-39933-0

[8] Pouyan, A., et al. (2025) Glioblastoma Multiforme: Insights into Pathogenesis, Key Signaling Pathways, and Therapeutic Strategies. *Molecular Cancer*, **24**, Article No. 58.

[9] Noushmehr, H., Weisenberger, D.J., Diefes, K., Phillips, H.S., Pujara, K., Berman, B.P., *et al.* (2010) Identification of a CPG Island Methylator Phenotype That Defines a Distinct Subgroup of Glioma. *Cancer Cell*, **17**, 510-522. https://doi.org/10.1016/j.ccr.2010.03.017

[10] Drexler, R., Khatri, R., Schüller, U., Eckhardt, A., Ryba, A., Sauvigny, T., *et al.* (2024) Temporal Change of DNA Methylation Subclasses between Matched Newly Diagnosed and Recurrent Glioblastoma. *Acta Neuropathologica*, **147**, Article No. 21. https://doi.org/10.1007/s00401-023-02677-8

[11] Babaei Rikan, S., Sorayaie Azar, A., Naemi, A., Bagherzadeh Mohasefi, J., Pirnejad, H. and Wiil, U.K. (2024) Survival Prediction of Glioblastoma Patients Using Modern Deep Learning and Machine Learning Techniques. *Scientific Reports*, **14**, Article No. 2371. https://doi.org/10.1038/s41598-024-53006-2

[12] Little, R.J.A. (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, **83**, 1198-1202. https://doi.org/10.1080/01621459.1988.10478722

[13] Keselman, H.J., Othman, A.R., Wilcox, R.R. and Fradette, K. (2004) The New and Improved Two-Sample *t* Test. *Psychological Science*, **15**, 47-51. https://doi.org/10.1111/j.0963-7214.2004.01501008.x

[14] Wang, H., Tang, J., Wu, M., Wang, X. and Zhang, T. (2022) Application of Machine Learning Missing Data Imputation Techniques in Clinical Decision Making: Taking the Discharge Assessment of Patients with Spontaneous Supratentorial Intracerebral Hemorrhage as an Example. *BMC Medical Informatics and Decision Making*, **22**, Article No. 13. https://doi.org/10.1186/s12911-022-01752-6

[15] Dong, X., Lin, L., Zhang, R., Zhao, Y., Christiani, D.C., Wei, Y., *et al.* (2018) TOBMI: Trans-Omics Block Missing Data Imputation Using a K-Nearest Neighbor Weighted Approach. *Bioinformatics*, **35**, 1278-1283. https://doi.org/10.1093/bioinformatics/bty796

[16] Al Mamlook, R.E., Nasayreh, A., Gharaibeh, H. and Shrestha, S. (2023) Classification of Cancer Genome Atlas Glioblastoma Multiform (TCGA-GBM) Using Machine Learning Method. 2023 *IEEE International Conference on Electro Information Technology* (*eIT*), Romeoville, 18-20 May 2023, 265-270. https://doi.org/10.1109/eit57321.2023.10187283

[17] Alin, A. (2010) Multicollinearity. *WIREs Computational Statistics*, **2**, 370-374. https://doi.org/10.1002/wics.84

[18] Mohammad-Djafari, A. (2021) Regularization, Bayesian Inference, and Machine Learning Methods for Inverse Problems. *Entropy*, **23**, Article 1673. https://doi.org/10.3390/e23121673

[19] Shekar, B.H. and Dagnew, G. (2019) Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. 2019 *Second International Conference on Advanced Computational and Communication Paradigms* (*ICACCP*), Gangtok, 25-28 February 2019, 1-8. https://doi.org/10.1109/icaccp.2019.8882943

[20] Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J.B., *et al.* (2005) Radiotherapy Plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *New England Journal of Medicine*, **352**, 987-996. https://doi.org/10.1056/nejmoa043330

[21] Azoulay, M., Santos, F., Souhami, L., Panet-Raymond, V., Petrecca, K., Owen, S., *et al.* (2015) Comparison of Radiation Regimens in the Treatment of Glioblastoma Multiforme: Results from a Single Institution. *Radiation Oncology*, **10**, Article No. 106. https://doi.org/10.1186/s13014-015-0396-6

[22] Jiang, C., Mogilevsky, C., Belal, Z., Kurtz, G. and Alonso-Basanta, M. (2023) Hypofractionation in Glioblastoma: An Overview of Palliative, Definitive, and Exploratory Uses. *Cancers*, **15**, Article 5650. https://doi.org/10.3390/cancers15235650