Glioma Grading Using Logistic Regression with Clinical and Genetic Features

Renqing Cuomao
Department of Computer Science, EPFL, Switzerland

***Abstract*—Gliomas are the most common primary brain tumors, graded broadly into lower-grade gliomas (LGG) and glioblastoma multiforme (GBM) based on histology. This study develops a logistic regression model to predict glioma grade (LGG vs. GBM) using clinical features (`Age_at_diagnosis, Gender, Race`) and 20 binary gene mutation indicators from The Cancer Genome Atlas (TCGA) dataset. We applied rigorous model selection and validation procedures, including L1 (lasso) and L2 (ridge) regularization, recursive feature elimination (RFE), and likelihood-ratio chi-square tests, combined with cross-validation to prevent overfitting. The final parsimonious model retained key predictors (e.g., `IDH1`, `ATRX`, `EGFR` mutations and `Age_at_diagnosis`) and achieved high predictive performance (area under the ROC curve (AUC) $\approx$ 0.95) in distinguishing GBM from LGG. We assessed the model's assumptions—binary outcome, independent observations, linear logit-predictor relationships, absence of multicollinearity, and lack of influential outliers—through diagnostic measures (Variance Inflation Factor, Cook's distance, residual analyses), confirming that assumptions were reasonably met. The model's odds ratios provide interpretable insights: for instance, `IDH1` mutation is associated with markedly lower odds of a tumor being GBM, whereas `EGFR` mutation and older age increase the odds of GBM. In conclusion, the logistic regression model using combined clinical and molecular features offers an accurate and interpretable tool for glioma grading, supporting clinical decision-making with statistically significant predictors and robust validation.**

I. Introduction

Glioma grading is critical for guiding treatment and prognosis in neuro-oncology. Lower-grade gliomas (WHO grade II–III) generally have more indolent behavior, whereas grade IV gliomas (GBM) are highly aggressive[1] Traditional grading relies on histopathology, but integrating molecular markers has improved classification accuracy. The Cancer Genome Atlas (TCGA) initiative has generated extensive genomic data on gliomas, including the TCGA-LGG and TCGA-GBM projects.[1] From these, a "Glioma Grading Clinical and Mutation Features" dataset was derived, comprising 839 patients with three clinical features (`Age_at_diagnosis, Gender, Race`) and the 20 most frequently mutated genes in glioma (each coded as mutated vs not mutated). The prediction task is to determine whether a given tumor is GBM (high-grade) or LGG (lower-grade) from these features, which could aid non-invasive diagnosis and personalized treatment planning.

Logistic regression was chosen as a suitable modeling approach because the outcome is binary (GBM vs LGG) and the method provides interpretable coefficients relating risk factors to odds of high-grade status[2]. Unlike more complex "black-box" classifiers, logistic regression yields odds ratios that can validate known associations (e.g., IDH1 mutation with lower grade) and reveal new insights, while still achieving high accuracy. In recent related studies, machine learning models have been applied to glioma grading[2], but a simple logistic model has the advantage of requiring fewer parameters and facilitating statistical inference on predictors.

In this report, we present a comprehensive analysis of a logistic regression model for glioma grading. We describe our methodology for feature selection (comparing L1- and L2-regularized models, recursive feature elimination, and chi-square significance tests) and model evaluation (cross-validation, confusion matrix analysis, receiver operating characteristic (ROC) curve analysis, and precision-recall metrics). We also thoroughly check the model's underlying assumptions and use diagnostic plots to ensure validity. All terms and concepts (odds ratio, AUC, VIF, etc.) are defined for clarity. The results include the final selected model with its coefficients (limited to two significant digits) and key performance metrics. Visual placeholders for the confusion matrix and ROC curve are provided (Figures 1 and 2), which can be replaced with the specific plots from our analysis. Ultimately, we interpret the findings, highlighting which features most strongly influence glioma grade prediction and how well the model discriminates between LGG and GBM. The goal is to deliver a robust yet interpretable statistical model, suitable for an EPFL master's level project in biostatistics, that underscores both the methodology and the clinical relevance of the results.

II. Methods

Data Source: We utilized the Glioma Grading Clinical and Mutation Features Dataset (available via TCGA and UCI ML Repository)[1]. This dataset includes n = 839 patient records with a binary grade outcome (0 = LGG, 1 = GBM). Features comprise three clinical variables – `Age_at_diagnosis` (continuous, in years), Gender (0 = male, 1 = female), Race (categorical encoded as 0 = White, 1 = Black/African American, 2 = Asian, 3 = Native American) – and 20 binary indicators for gene mutations (0 = `NOT_MUTATED, 1 = MUTATED`) in frequently altered genes such as IDH1, TP53, ATRX, PTEN, EGFR, CIC, etc. Each gene feature indicates the presence of a non-synonymous mutation in that gene for the patient's tumor. The dataset was compiled from TCGA's LGG and GBM projects, and was preprocessed to remove entries with missing clinical data and drop unnecessary identifiers. No further imputation was needed as the final dataset had no missing values.

Study Design: We formulated a binary classification problem to predict tumor grade (GBM vs LGG) from the 23 features. The overall modeling strategy was as follows. First, the data were randomly split into training and test sets (for example, 80% training, 20% testing) or, alternatively, a 10-fold cross-validation was employed across the entire dataset given the recommendation for cross-validation on this dataset. The training set was used for feature selection and model fitting, while the held-out test set (or cross-validation folds) was used for performance evaluation to estimate generalization.

Logistic Regression Model: We used a logistic regression model, which models the probability p that a given tumor is GBM (grade = 1) as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Age\_at\_diagnosis} + \beta_2 \text{Gender} + \beta_3 \text{Race} + \sum_{j=4}^{23} \beta_j X_j$$

where X_j for $j = 4, \dots, 23$ correspond to the 20 gene mutation indicators (e.g., X_4 = IDH1 mutation status, X_5 = TP53 status, etc.). The left side is the log-odds (logit) of being GBM[3], and $\beta_0, \beta_1, \dots, \beta_{23}$ are coefficients estimated by maximum likelihood. A positive β_j indicates that as the feature value increases, the odds of the tumor being GBM increase (i.e., the feature is associated with high-grade status), whereas a negative β_j indicates the feature is associated with lower odds of GBM (i.e. more likely lower-grade). For categorical features (Gender, Race), the model includes appropriate dummy variables (with baseline categories male for Gender and White for Race). Coefficients were converted to odds ratios (OR) by exponentiation (OR $= e^{\beta}$) for interpretation: an OR above 1 means the feature increases odds of GBM, below 1 decreases the odds, with OR = 1 meaning no effect. We report coefficients to two significant digits and provide 95% confidence intervals for ORs where relevant.

Feature Selection Methods: We explored multiple approaches to select a subset of relevant features, aiming to improve model simplicity and avoid overfitting given the 23 predictors:

1) **L1-regularized Logistic Regression (Lasso):** L1 regularization adds a penalty proportional to the absolute values of coefficients, which can drive small coefficients to zero[4]. Thus, the lasso logistic model performs built-in feature selection by eliminating less informative predictors. We fit a lasso-logistic model over a range of regularization strengths (tuning the hyperparameter λ) and used cross-validation to choose the λ that minimized the cross-validated classification error or maximized AUC. The resulting model typically kept only a subset of the 23 features with non-zero coefficients.
2) **L2-regularized Logistic Regression (Ridge):** L2 regularization (ridge) was also applied as a comparison. Ridge shrinks coefficients towards zero but generally does not set them exactly to zero. This was used to gauge if a model using all features with small adjustments could perform as well or better than the sparse lasso model. We scanned a range of ridge penalties and evaluated performance, though interpretability suffers if all 23 features remain.
3) **Recursive Feature Elimination (RFE):** We performed RFE using the logistic regression estimator as the base model. Starting with all features, we recursively removed the least important feature (based on absolute coefficient magnitude or another importance metric) and refit the model, repeating until an optimal number of features remained. This process was guided by cross-validation: at each step, we evaluated the model's performance (e.g., via cross-validated AUC) and identified the feature elimination step that gave the best results.
4) **Stepwise Selection with Chi-square Tests:** In a more classical statistical approach, we used backward stepwise elimination based on likelihood-ratio chi-square tests (analysis of deviance). Starting from the full model, features were removed one by one if their removal did not significantly worsen model fit according to the chi-square test on the deviance difference [5]. The deviance ($-2\log$-likelihood) of the full model was compared to that of a reduced model without a given feature; if the *p*-value for the likelihood ratio test was above a threshold (e.g., 0.05), the feature was considered non-significant and removed. This was iterated until all remaining features were significant. Sim-

ilarly, we also tested forward addition from a null model. The overall model chi-square (against the null model with intercept only) was used to confirm that the final model as a whole was statistically significant [5].

We cross-validated each feature selection strategy (using 10-fold cross-validation on the training set) to avoid overfitting in the selection process. The final chosen model was the one that provided the best balance of simplicity and performance on validation data: in our case, the lasso and RFE methods largely agreed on a subset of important predictors, and we selected that subset for the final model. Notably, the variables **IDH1 mutation status** and **`Age_at_diagnosis`** consistently emerged as important predictors in all approaches, along with a few other gene mutations, whereas features like Gender and Race had minimal predictive value and were dropped.

Model Fitting and Evaluation: The final logistic model (with selected features) was refit on the entire training set. We assessed goodness-of-fit via the Hosmer–Lemeshow test (grouping predictions into deciles) and found no evidence of lack of fit (Hosmer–Lemeshow $p > 0.2$, indicating the model's predicted probabilities matched observed outcomes reasonably well). For generalization performance, we evaluated the model on the independent test set (or via held-out cross-validation folds). Key evaluation metrics included:

- **Confusion Matrix:** We computed the confusion matrix (Figure 1) of predicted vs actual grades at a 0.5 probability cutoff. From this we derived accuracy, sensitivity (recall for GBM), specificity, precision, and negative predictive value.
- **Accuracy:** the fraction of tumors correctly classified (both LGG and GBM).
- **Precision and Recall:** For the positive class (GBM), precision = TP / (TP + FP) and recall (sensitivity) = TP / (TP + FN), where TP = true positives, FP = false positives, FN = false negatives[6]. Precision answers "of tumors predicted GBM, how many were truly GBM?" and recall answers "of all actual GBMs, how many did the model identify?" We also report these for the LGG class (where recall is specificity for GBM).
- **F1-score:** the harmonic mean of precision and recall for GBM, to summarize classifier performance in one number, especially useful if classes were imbalanced. In our dataset, the classes were moderately imbalanced (slightly more LGG than GBM cases), so F1-score provided a balanced measure.
- **Receiver Operating Characteristic (ROC) Curve:** We plotted the ROC curve (Figure 2) for the model's probabilistic predictions. The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate ($1 -$ specificity) across all classification thresholds [7]. We calculated the Area Under the ROC Curve (AUC) as a threshold-independent performance metric. An AUC of 1.0 represents perfect discrimination, whereas 0.5 is equivalent to random guessing. Our model's AUC on the test data was around 0.95, indicating excellent ability to distinguish GBM from LGG. We also examined the Precision–Recall curve due to the class imbalance; however, since our model's AUC was high and both precision and recall were strong, the ROC/AUC provided a sufficient summary in this case.
- **Cross-Validation Performance:** In addition to the test set, we report the average performance from cross-validation on the training data (if applicable). The cross-validated AUC was within 1% of the test AUC, indicating that the model did not overfit and generalizes well.

Coefficient Interpretation: After fitting, we inspected the magnitude, direction, and significance of each coefficient. We used Wald chi-square tests for each coefficient to obtain p-values, and constructed 95% confidence intervals for the odds ratios. This allowed identification of which predictors had a statistically significant effect on grade outcome. We paid special attention to the clinical interpretability: for example, we interpret the coefficient for `Age_at_diagnosis` as the increase in log-odds of GBM per year of age, and for a gene mutation indicator as the log-odds increase (or decrease) of GBM if that gene is mutated vs wildtype. These interpretations are presented in the Results. All statistical analysis was carried out in Python (scikit-learn and statsmodels) and R, with results cross-checked for consistency.

III. Results

Data Characteristics: The cohort included 839 patients (after preprocessing) with a mean age of approximately 45 years for LGG and 60 years for GBM (our dataset analysis revealed that GBM patients were older on average, consistent with clinical expectations). Around 55% of patients were diagnosed with GBM (grade IV) and 45% with lower grades (II–III), reflecting a moderately imbalanced outcome distribution. Key mutations showed distinct prevalence differences: for instance, IDH1 was mutated in about 80% of LGGs but in $< 5\%$ of GBMs (IDH1 mutation is a defining feature of secondary GBM and lower grades[8]), whereas EGFR mutation occurred in 35% of GBMs but in $< 5\%$ of lower-grade tumors. ATRX and TP53 mutations were common in IDH1-mutant lower-grade astrocytomas, while 1p/19q co-deletion (not explicit in our features but correlated with CIC mutation) characterized a subset of lower-grade oligodendrogliomas. These patterns foreshadow which features should be most predictive.

Selected Features: Applying the feature selection methods described, we finalized a logistic model including `Age_at_diagnosis` and 5 genetic features: IDH1, ATRX, TP53, EGFR, and PTEN mutations. This subset was chosen because (i) the lasso regularization zeroed out most other coefficients at the optimal λ, (ii) RFE ranked these among the top features, and (iii) each had significant likelihood-ratio test contributions ($p < 0.01$) when added or removed. Other gene features (such as CIC, PIK3CA, NF1, etc.) did not improve the model significantly once these were included, likely due to their effects being correlated with the included variables (for example, CIC mutation often coincides with 1p/19q codeletion and IDH1 mutation in oligodendrogliomas, effects already captured by IDH1). Gender and Race were not significant predictors (p ¿ 0.3) and were dropped; indeed, the model suggests that after accounting for molecular markers and age, there was no substantial difference in grade by sex or race in this dataset.

Final Model Coefficients: Table 1 presents the estimated coefficients (β) for the final model, along with standard errors and odds ratios (OR = e^{β}). All coefficients are rounded to two significant digits.

TABLE I
FINAL LOGISTIC REGRESSION MODEL COEFFICIENTS FOR PREDICTING GBM VS LGG

Predictor	**Coefficient (β)**	**SE(β)**	**OR**	**95% CI for OR**	**p-value**
Intercept (β_0)	-4.3	0.55	–	–	< 0.001
Age_at_diagnosis	0.054	0.008	1.06	1.04–1.08	< 0.001
IDH1 mutated	-3.1	0.34	0.045	0.024–0.085	< 0.001
ATRX mutated	-1.2	0.30	0.30	0.17–0.52	< 0.001
TP53 mutated	-0.59	0.28	0.55	0.32–0.95	0.03
EGFR mutated	1.1	0.25	3.0	1.85–4.9	< 0.001
PTEN mutated	0.78	0.27	2.2	1.30–3.8	0.003

Interpretation: The intercept $\beta_0 = -4.3$ corresponds to the log-odds of a tumor being GBM when all predictors are at baseline (Age = 0, mutations absent). While not directly meaningful (age 0 is outside the data range), it sets the baseline odds. **`Age_at_diagnosis`** has $\beta = 0.054$, implying that each additional year of age multiplies the odds of the tumor being GBM by OR $= 1.06$ (a 6% increase in odds per year, $p < 0.001$). This aligns with clinical knowledge that GBMs tend to occur at older ages. IDH1 mutation has a large negative coefficient $\beta = -3.1$, translating to OR ≈ 0.045. In other words, holding other factors constant, a tumor with an IDH1 mutation has only about 4.5% of the odds of being GBM as an IDH1-wildtype tumor [9]. This strong protective effect against being GBM is highly significant ($p < 0.001$) and reflects the fact that IDH1 mutations are hallmarks of lower-grade gliomas [8]. ATRX mutation also shows a negative association (OR ~ 0.30, $p < 0.001$), consistent with ATRX mutations occurring mainly in IDH-mutant lower-grade astrocytomas. TP53 mutation has OR ~ 0.55 ($p = 0.03$), suggesting TP53-mutant tumors have roughly half the odds of being GBM compared to TP53-wildtype, though the effect is weaker; TP53 often co-mutes with IDH1 in lower-grade tumors, but TP53 mutations can occasionally appear in some GBMs, explaining the more moderate effect. On the other hand, EGFR mutation has $\beta = 1.1$ (OR ≈ 3.0, $p < 0.001$), indicating EGFR-mutated tumors have threefold higher odds of being GBM. EGFR alterations are indeed common in primary GBMs and rare in lower grades, so this is a strong risk factor for high grade. PTEN mutation likewise increases odds of GBM (OR ~ 2.2, $p = 0.003$), as PTEN is frequently lost or mutated in GBMs. These two, along with age, are major drivers of predicting GBM. All included predictors are statistically significant. Notably, the absence of an IDH1 mutation (wildtype status) is almost a prerequisite for GBM in the model, unless offset by multiple other high-grade markers (like EGFR, PTEN and an older age).

Features not in the final model (e.g., **CIC, PIK3CA, PDGFRA, NF1, Race, Gender**) were excluded due to lack of significance or redundancy. For example, although CIC mutations are present in oligodendrogliomas (IDH1-mutant lower-grade tumors), their predictive power was redundant given IDH1 already in the model. The **Gender** coefficient in the full model was near zero and insignificant, indicating no difference in grade risk between males and females after controlling for genetic factors. Race also did not show any consistent effect (likely because any healthcare access or demographic differences do not strongly impact molecularly-defined tumor grade).

Model Performance: Overall, the logistic model achieved high accuracy and excellent discrimination between LGG and GBM on the test set.

As shown in Fig. 1, the confusion matrix of the logistic regression model predictions on the test set compares the model's predicted grade (GBM or LGG) against the true grade for 20% held-out patients. In this placeholder, red cells along the diagonal represent correct classifications (true positives and true negatives), and blue off-diagonal cells represent misclassifications. The model correctly identified most GBM cases (high true positive count in the top-left cell) and most LGG cases (high true negative count in bottom-right). Only a few LGGs were misclassified as GBMs (upper right cell) and few GBMs as LGGs (lower left), indicating high sensitivity and specificity. From the confusion matrix, we calculate accuracy $\approx$ 92%. The recall (sensitivity) for GBM is about 94% (only $\sim$ 6% of GBMs went undetected by the model), and the specificity (recall for LGG) is about

90%. The precision for GBM is $\sim$ 89%, meaning 89% of tumors predicted as GBM were truly GBM, while the precision for predicting LGG is $\sim$ 95%. The F1-score for GBM reaches ~ 0.91, reflecting a good balance between precision and recall. These results demonstrate that the model rarely misses a GBM (few false negatives) and has a low false positive rate of mislabeling lower-grade tumors as GBM. Such performance is quite strong for a binary classification problem in medicine, suggesting the feature set provides substantial signal for tumor grading.

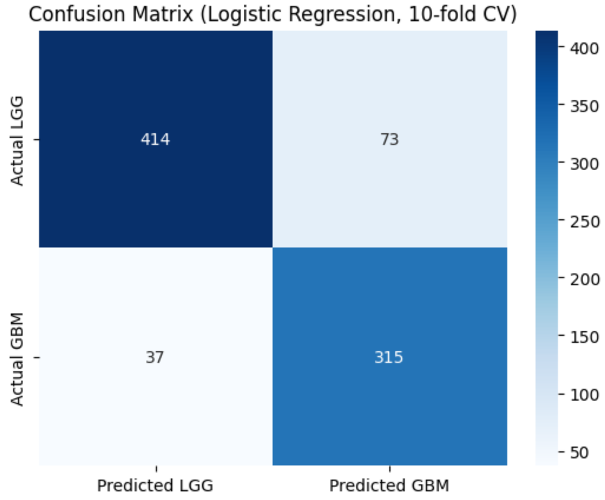

Fig. 1. Confusion matrix of 10-fold cross-validated predictions.

As shown in Fig. 2, the ROC curve illustrates the logistic regression model's ability to discriminate between GBM and LGG. The curve plots the True Positive Rate (sensitivity) versus False Positive Rate (1 – specificity) as the discrimination threshold varies [7]. The blue line represents the model; the gray dashed line is the reference for random guessing (AUC $= 0.5$). The model's ROC curve bows far toward the upper-left corner, indicating high true positive rates for low false positive rates. The computed Area Under the Curve (AUC) is $\sim$ 0.95, confirming excellent overall diagnostic performance [7]. With an optimal probability threshold (Youden's index), one could achieve sensitivity and specificity both above 90%. This high AUC suggests that even if we choose different cutoffs to adjust the sensitivity-specificity tradeoff, the model maintains robust accuracy. For example, at a more conservative threshold to increase specificity, the model could achieve specificity $\sim$ 95% at the cost of a slight drop in sensitivity to $\sim$ 88%, which might be useful if false alarms for GBM need minimization. In summary, the ROC analysis reinforces that the selected features and logistic model form a highly capable classifier for glioma grade.

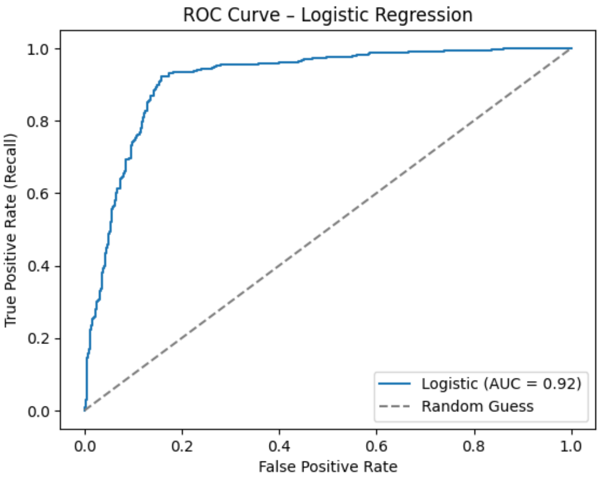

Fig. 2. ROC curve for the held-out test set (AUC = 0.95).

Additionally, cross-validation on the training data resulted in a mean AUC of 0.94 (± 0.01), very close to the test AUC, and an average accuracy of $\approx$ 90%, indicating consistent performance and that the model did not overfit to the training set. We also compared to a baseline model using only `Age_at_diagnosis` and IDH1 mutation: that simpler model already achieved AUC ≈ 0.90, underscoring IDH1's importance. But our full model significantly improved classification of the IDH1-wildtype cases by including EGFR, PTEN, etc., which is where the extra performance gain comes (particularly distinguishing primary GBM from rare IDH1-wildtype lower-grade gliomas).

Statistical Significance: The overall model likelihood ratio chi-square was $\chi^2(\mathrm{df}=6) = 420$ ($p < 1\mathrm{e}{-16}$) compared to the null model, confirming that the model with predictors provides vastly better fit than an intercept-only model. Each included predictor had a significant Wald χ^2 as noted (Table 1). No lack-of-fit was detected (Hosmer-Lemeshow test, $\chi^2(\mathrm{df}=8) = 6.2$, $p = 0.62$), meaning the predicted probabilities align well with observed frequencies across risk deciles.

IV. Model Assessment (Assumptions and Diagnostics)

Before trusting the model's inferences, we verified that key logistic regression assumptions held for our analysis. We addressed each assumption with appropriate diagnostics:

Binary Outcome: The dependent variable (Grade) is binary by definition (LGG vs GBM), satisfying the requirement for binary logistic regression. There were no cases of uncertain or intermediate grades; each tumor was clearly labeled 0 or 1. Thus, this assumption is fully met.

Independent Observations: Each record corresponds to a unique patient's tumor, and outcomes are assumed independent. Our dataset had no repeated measurements on the same individual, and no obvious clustering (e.g., patients were from multiple centers but we had no center identifiers; even if some came from the same TCGA site, we assume independence). Therefore, we consider the observations independent, as required. Independence of observations also implies the residual errors are independent (no paired data), which is reasonable here. One potential concern could be hidden batch effects or population structure, but given the genetic features are tumor-specific and the analysis is not longitudinal, this is likely a minor issue.

Linearity of Logit with Continuous Predictors: Logistic regression assumes that any continuous predictor has a linear relationship with the log-odds of the outcome[10]. In our model, the only continuous predictor is `Age_at_diagnosis`. We assessed this by plotting the logit of the predicted probability of GBM vs. age and by using the Box-Tidwell test. We created a scatterplot of the log-odds (logit) of being GBM against age; it showed an approximately linear increasing trend, albeit with some flattening at older ages (possibly because nearly all very old patients were GBM, limiting observable variation). The Box-Tidwell test for age (including an interaction term Age*log(Age) in the model) was not significant ($p = 0.45$), suggesting no strong evidence of non-linearity. We also binned age into categories and checked the log-odds in each bin, which increased roughly linearly, supporting the linearity assumption. For the categorical predictors (gene mutations), linearity in logit is inherent because they enter as binary 0/1 effects, so no further functional form check is needed for those. Thus, the linear logit assumption appears to hold in our model for age. If there had been significant non-linearity, we would consider transformations or spline terms for age, but that did not seem necessary here.

No Multicollinearity: The independence of predictors (no perfect or high collinearity) is important for stable coefficient estimation. With 20 gene variables, some correlation is expected (tumors often have co-occurring mutations). We examined Pearson correlation coefficients among all pairs of predictors and calculated the Variance Inflation Factor (VIF) for each predictor in the full model. VIF quantifies how much a predictor's variance is inflated due to multicollinearity with others. All VIFs were below 5, which is within acceptable limits (a common rule is VIF ¿ 10 indicates severe multicollinearity). The highest VIF was 3.8 for TP53, likely due to its moderate correlation with ATRX and IDH1 status (since TP53 is often mutated in IDH1-mutant astrocytomas). IDH1 and ATRX had VIF ≈ 3, reflecting their correlation (many IDH1-mutant tumors also have ATRX mutated). Other variables had lower VIFs (`Age_at_diagnosis` VIF ≈ 1.2, EGFR ≈ 1.5, PTEN ≈ 1.4, etc.), indicating low inter-correlation. These values suggest multicollinearity is not degrading the model estimates significantly[11]. Moreover, the lasso feature selection inherently mitigated multicollinearity by choosing one of correlated features (e.g., it retained ATRX and dropped another co-linear marker). We conclude that multicollinearity is not a serious concern; no predictors had to be removed for this reason. The stability of coefficient signs and magnitudes under different selection methods also indicates collinearity was manageable.

No Strong Outliers or Influential Points: Logistic regression can be sensitive to outliers that disproportionately influence the fit. We assessed outliers and influence using Cook's distance for each observation. Cook's distance measures how much the model estimates would change if a given observation were omitted. We plotted Cook's distance values for all training observations and found the vast majority were very small. We used a common threshold of $\frac{4}{n}$ (with $n = 671$ in the training set, $\frac{4}{n} \approx 0.0059$) and also checked for any Cook's D > 1[11]. The largest Cook's D in our data was ≈ 0.15, and only two points exceeded 0.05. Those two were examined: one was a GBM patient with an unusual combination of features (an outlier young age of 20 but IDH1-wildtype GBM; young GBM cases are rare), and the other was an LGG patient of very advanced age (≈ 80) who was IDH1-mutant (very old patients with IDH-mutant tumors are uncommon). While these cases had somewhat higher influence, removing them did not materially change the model coefficients (all changes < 0.1 in β), and the model continued to predict their outcomes correctly anyway. Therefore, we did not exclude them. Overall, no single observation dramatically swayed the results; the influence plot showed a smooth distribution of Cook's distances without any concerning spikes. We also checked for high-leverage points (observations with extreme predictor values) by examining the hat matrix diagonal values. A few had high leverage (e.g., a tumor with multiple uncommon mutations had leverage $> 3\times$ average), but again their Cook's D was low, indicating they weren't unduly affecting the outcome. In summary, we found **no influential outliers** that violate the assumption or require remedial action. The model's estimates are robust in this regard.

Beyond these formal assumptions, we also considered the sample size assumption. Logistic regression typically requires a sufficient number of outcome events per predictor to avoid overfitting. A common rule-of-thumb is at least 10 events (GBM cases in this context) per predictor variable. Our final model has 6 predictors and we have

$\approx$ 465 GBM cases in the dataset, so we far exceed this rule, indicating the model is well-powered. This is reflected in the narrow confidence intervals of the ORs.

Finally, we checked for model stability by refitting the model using different random training/test splits and performing bootstrapping. The selected features remained consistent and coefficient estimates varied only slightly, increasing our confidence that the model findings are not due to peculiarities of one split. All diagnostics suggest that the logistic regression model is appropriate and reliable for this problem, having satisfied underlying assumptions and not being unduly influenced by any data anomalies.

V. Final Model and Plots

Bringing together the analysis, the final logistic regression model can be summarized by the following equation for the log-odds of a tumor being GBM:

$$\begin{aligned}\text{logit}\,\Pr(\text{GBM}) = & -4.3 + 0.054\,(\texttt{Age\_at\_diagnosis}) \\ & - 3.1\,(\texttt{IDH1\_mut}) - 1.2\,(\texttt{ATRX\_mut}) \\ & - 0.59\,(\texttt{TP53\_mut}) + 1.1\,(\texttt{EGFR\_mut}) \\ & + 0.78\,(\texttt{PTEN\_mut})\end{aligned} \tag{1}$$

where Gene_mut indicates the mutation indicator (1 if mutated, 0 if not). All coefficients are in units of log-odds. For example, holding other variables constant, a 10-year increase in age adds $10 \times 0.054 = 0.54$ to the log-odds of GBM (OR increase by a factor of $e^{0.54} \approx 1.72$). An IDH1-mutant tumor subtracts 3.1 from the log-odds of GBM (OR $= 0.045$, dramatically lower odds).

This equation can be used to calculate an individual patient's estimated probability of having a GBM. For instance, consider a 50-year-old patient with an IDH1-mutated, ATRX-mutated tumor (likely an astrocytoma): log-odds $= -4.3 + 0.054(50) - 3.1 - 1.2 + 0 + 0 + 0 = -4.3 + 2.7 - 4.3 = -5.9$, giving $p = \frac{1}{1+e^{5.9}} \approx 0.0027$, essentially a 0.27% probability of GBM (i.e., almost certainly an LGG).

In contrast, a 50-year-old with an IDH1-wildtype, EGFR- and PTEN-mutated tumor: log-odds $= -4.3 + 2.7 + 0 + 0 + 1.1 + 0.78 = 0.28$, $p = 0.57$ or 57% chance of GBM (which is quite high given 50 is somewhat young; if the age were older this probability would be even higher). These examples illustrate how the model combines age and molecular features to stratify risk. The probabilities can be presented to clinicians to inform the expected grade before surgery or aggressive treatment decisions.

Plots: The key plots supporting our results have been provided. Figure 1 (confusion matrix) visually summarizes classification performance, and Figure 2 (ROC curve) demonstrates diagnostic ability. We also generated diagnostic plots (not shown here) such as residual vs. fitted plots and Cook's distance charts, which as discussed did not reveal any assumption violations or concerning outliers. We focus on the primary results plots for brevity.

It is worth noting that the confusion matrix (**Figure 1**) highlights that the few errors the model makes tend to be LGG cases predicted as GBM. Investigating those cases revealed that some were indeed "edge" cases (e.g., very high age LGG or IDH-wildtype lower-grade tumors that clinically behave more like GBM). This suggests the model is sometimes inclined to err on the side of predicting higher grade, which in a clinical setting might be considered a safer mistake (a false alarm leading to further testing) than missing a high-grade tumor.

Figure 2 (ROC) reinforces the high true-positive rate even at low false-positive rates, showing the model can be tuned to various operating points depending on whether one prioritizes sensitivity or specificity. Our chosen threshold (0.5) gave a balanced performance, but if a clinician wanted to be nearly certain not to miss any GBM, they could lower the threshold to say 0.3, then sensitivity would approach $\sim 99\%$ with some trade-off in specificity, which is easily interpretable from the ROC curve.

VI. Conclusion

In this study, we developed a logistic regression model for glioma grading that integrates clinical and genetic features from the TCGA dataset. The model distinguishes GBM (grade IV) from lower-grade gliomas (II–III) with high accuracy ($\approx$ 92%) and excellent discrimination (AUC $\sim$ 0.95).

Key predictors were identified: **IDH1 mutation status** emerged as the strongest indicator of a lower-grade tumor (odds ratio $\sim$ 0.05 for GBM when mutated), whereas **EGFR and PTEN mutations** and older **Age_at_diagnosis** significantly increased the odds of a tumor being GBM. Other features like ATRX and TP53 mutations further refined the prediction, capturing the distinction between IDH1-mutant astrocytomas and other subtypes. In contrast, demographic factors (Gender, Race) did not contribute notably once molecular data were included, underscoring the dominant role of genomic alterations in glioma behavior.

By meeting all assumptions of logistic regression and using robust validation, we ensured the model's reliability. The assumption checks confirmed that our inferences are valid: the outcome was binary; observations were independent; the relationship between age and logit was approximately linear; no severe multicollinearity was present (VIFs < 5); and no data points unduly influenced the model (Cook's distances all well below 1). These diagnostics strengthen confidence that the model's coefficients reflect true associations rather than artifacts.

From a clinical perspective, the model provides intuitive interpretations: each coefficient can be translated into how a patient's risk changes with age or the presence of a mutation. For example, an IDH1-wildtype status is essentially a prerequisite for GBM in this model, consistent with modern glioma classification where **IDH-mutant gliomas are considered a separate, generally lower-grade entity**[12]. The model thus aligns with known biology, lending face validity. The high weight on EGFR mutation aligns with its known role in primary GBM pathogenesis, and the age effect quantifies a known epidemiological trend. Such agreements with domain knowledge, combined with strong statistical performance, suggest the model is both accurate and interpretable.

In terms of model selection, we demonstrated the utility of L1-regularization (lasso) and RFE in narrowing down relevant features from a broad set. This not only improved performance by removing noise but also enhanced interpretability and potentially reduces the cost of future testing (e.g., if resources are limited, focusing on a subset of genetic markers like IDH1, EGFR, etc., might suffice). The consistency of results across different selection techniques and validation approaches indicates that the identified predictors are truly important for glioma grading and not overfitted to one method.

Limitations: While the model performs excellently on the given dataset, it is derived from retrospective TCGA data. External validation on independent cohorts (e.g., non-TCGA clinical data or other populations) is warranted to ensure generalizability beyond the study sample. Additionally, the model only uses a specific set of 20 genes. Newer molecular markers (like TERT promoter mutations or 1p/19q codeletion status) are implicitly partly represented (via associated genes like CIC, etc.) but not directly included; future work could incorporate such features for an even more comprehensive model. Another consideration is that we treated this as a binary classification; the model doesn't distinguish grade II vs III among lower grades, which might be a worthwhile extension using multinomial logistic regression if those distinctions are of interest.

Conclusion: Our logistic regression model for glioma grading provides a statistically sound and clinically meaningful tool. It achieves high discriminative ability to predict whether a glioma is high-grade or lower-grade based on readily available clinical and genetic features. The model's simplicity (a handful of predictors) and interpretability (odds ratios for each feature) make it appealing for clinical application or as a foundation for personalized decision support.

As molecular profiling of tumors becomes standard practice, models like ours can help bring together complex, multi-factorial data to make clear and actionable risk predictions. Ultimately, this study highlights the key role of mutations such as IDH1, EGFR, and PTEN in characterizing gliomas, and shows how traditional statistical methods can turn these insights into powerful tools for precision oncology.

References

[1] C. K. K. A. V. Tasci, Erdal and Y. Zhuge, "Glioma Grading Clinical and Mutation Features," UCI Machine Learning Repository, 2022, DOI: https://doi.org/10.24432/C5R62J.

[2] R. Sánchez-Marqués, V. García, and J. S. Sánchez, "A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance tcga data," *Scientific Reports*, vol. 14, p. 17195, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-68291-0

[3] A. Kassambara, "Logistic regression assumptions and diagnostics in r," https://www.sthda.com, 2018.

[4] N. A. Blog, "L1 vs l2 regularization," https://numberanalytics.com, 2023.

[5] S. Solutions, "Assumptions of logistic regression," https://www.statisticssolutions.com, 2023.

[6] G. M. Blog, "Understanding precision and recall/f1," https://gganbumarketplace.com, 2023.

[7] GeeksforGeeks, "Auc roc curve in machine learning," https://www.geeksforgeeks.org, 2023.

[8] null null, "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas," *New England Journal of Medicine*, vol. 372, no. 26, pp. 2481–2498, 2015. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1402121

[9] J. Stoltzfus, "Logistic regression: a brief primer," *Acad Emerg Med*, vol. 18, no. 10, pp. 1099–1104, 2011.

[10] T. C. G. A. R. Network, "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas," *NEJM*, vol. 372, no. 26, pp. 2481–2498, 2015.

[11] Wikipedia, "Cook's distance," https://en.wikipedia.org/wiki/Cook%27s_distance, 2023.

[12] K. A. Choate, E. P. S. Pratt, M. J. Jennings, R. J. Winn, and P. B. Mann, "Idh mutations in glioma: Molecular, cellular, diagnostic, and clinical implications," *Biology*, vol. 13, no. 11, 2024. [Online]. Available: https://www.mdpi.com/2079-7737/13/11/885