**Galina Glousker**

Senior Scientist, EPFL

galina.glousker@epfl.ch

**Data Science Project**

**The impact of molecular profiling of gliomas on diagnosis and prognosis**

**Final Project Report**

**07 July 2025**

# Abstract

Gliomas are a diverse group of primary brain tumors originating from glial cells (astrocytes, oligodendrocytes, and ependymal cells) in the central nervous system. Starting from 2016 WHO classification of CNS tumors and its revisions shifted the focus from purely histopathological analysis to an integrated diagnosis incorporating molecular markers. By integrating molecular markers into the diagnostic framework, clinicians can better stratify patients, tailor treatment plans, and predict clinical outcomes. This shift towards personalized medicine, based on well-established molecular data, is improving the management and treatment of gliomas, particularly in identifying patients who may benefit from specific therapies or have better prognoses.

In my project I will use genomic data from more than 1000 glioma patients from the study published in Clinical Cancer Research journal in 2019 [1] and apply supervised and unsupervised machine learning algorithms to identify relationships between genomic alterations and disease progression.

# Table of Contents

[Abstract 0](#_Toc202461962)

[Table of Contents 1](#_Toc202461963)

[**1. Project Objectives** 2](#_Toc202461964)

[**2. Methods** 3](#_Toc202461965)

[**3. Data** 3](#_Toc202461966)

[**4. Metadata** 4](#_Toc202461967)

[**5. Data Quality** 4](#_Toc202461968)

[**6. Data Flow** 4](#_Toc202461969)

[**7. Data Model** 5](#_Toc202461973)

[**8. Documentation** 5](#_Toc202461974)

[**9. Risks** 6](#_Toc202461975)

[**10. Exploratory data analysis** 6](#_Toc202461976)

**11. Machine learning analysis**  11

11.1 Clustering 11

11.1.1 Hierarchical clustering 11

11.1.2 Non-hierarchical clustering 12

11.2 Diagnostics 13

11.2.1 Tumor grading 13

11.3 Survival 14

**12. Discussion** 16

**13. Conclusions** 18

Statement 18

[References and Bibliography 19](#_Toc202461979)

# 

# **1. Project Objectives**

Malignant primary brain tumors remain among the most difficult cancers to treat, with a 5 year overall survival no greater than 35%. The most common malignant primary brain tumors in adults are gliomas. They include: astrocytomas, which arise from astrocytes; oligodendrogliomas, originating from myelin producing cells; and ependymomas, arising from ependymal cells (lining of brain ventricles). Gliomas are graded by the WHO classification into low grade gliomas (grade 1-2, least aggressive) and high grade gliomas (grade 3-4, most aggressive).

Central nervous system (CNS) tumors have long been classified based on histological findings supported by tissue-based tests (eg, immunohistochemical, ultrastructural). More recently, molecular biomarkers have gained importance in providing both ancillary and defining diagnostic information. For instance, tumors previously classified as “glioblastomas” may now be identified as IDH-mutant gliomas with a better prognosis. Genomic markers also provide prognostic information, with certain mutations (e.g., IDH mutations, MGMT promoter methylation) correlating with better survival, while others (e.g., EGFR amplification, H3 K27M mutation) indicating a worse prognosis. Genomic profiling can also identify potential targets for therapy, such as EGFR inhibitors for tumors with EGFR amplification and IDH inhibitors for IDH-mutant gliomas. (WHO classification [2, 3]).



Figure 1. Decision tree of glioma diagnostics based on WHO classification

The goal of my project is to analyze genomic data from a large cohort of adult patients with glioma to identify genomic alterations associated with clinical behavior, and use machine learning methods to be able to predict glioma outcomes in adult patients depending on their tumor molecular profile.

# **2. Methods**

**2.1 Infrastructure**

I local Python installations (Visual Studio Code, Anaconda Distribution for Python) on my personal computer.

**2.2 Software libraries and tools**

First data evaluation and some plots generation will be done using tools on cBioportal.org.

Project is done in Python using the following libraries for the analysis:

* Data handling & stats: math, numpy, pandas, scipy.
* Visualization: matplotlib, seaborn
* Classical stats: statsmodels
* Preprocessing & feature engineering: sklearn.preprocessing, sklearn.decomposition
* Clustering: sklearn.cluster
* Supervised learning: sklearn.linear\_model, sklearn.ensemble, sklearn.svm, sklearn.neighbors, GaussianMixture
* Survival analysis: lifelines, sksurv.util, Surv, sksurv.ensemble, sksurv.metrics
* Model selection & evaluation: sklearn.model\_selection, sklearn.metrics, sklearn.inspection, sklearn.feature\_selection, sklearn.utils.class\_weight

**2.3 Modeling, algorithms and statistical methods**

For exploring my dataset, I will use some descriptive statistics (to identify patients characteristics and demography, as well as most frequently mutated genes and distribution of diagnoses) and some inferential statistics (confidence intervals for survival, chi-square tests, non-parametic tests, survival analysis and regression models). To explore and model the relationships between patients’ molecular profiles and clinical outcomes I will start with unsupervised (PCA, UMAP, hierarchical clustering) and later supervised (linear regression, random forest and other) machine learning algorithms. I use 80:20 split and 5-fold cross-validation to train and validate the supervised ML models.

To validate accuracy of my classifications I use precision, recall, F1 and AUC scores metrics, to estimate how good my clustering analysis is I use Silhouette score, and for assessing quality of survival analysis I use C-index.

# **3. Data**

**3.1 Data collection and acquisition**

I use genomic data from more than 1000 glioma patients from the study published in Clinical Cancer Research in 2019. The chromosomal rearrangements, methylation and somatic mutations data are anonymized and publicly available from cBioportal.org (www.cbioportal.org/study?id ¼ glioma\_mskcc\_2019).

I use three data frames for my project. “Data\_mutations\_short” describes variants (mutations) in each gene of each patient (tumor\_sample\_barcode column) and variant classification (depending on how each mutation affects protein function). “Pathways” associate each gene with one of ten selected metabolic pathways (or none) to create a new feature. Then I use tumor\_sample\_barcode column to map each mutated gene and pathway to the patient in “clinical\_data” data frame that contains all the information about patients and that will be used for further analysis. Unfortunately, since the dataset contains unmatched Patient and Sample IDs, I had to omit almost 300 patients out of more than 1000, into my final data frame “clinical mutations\_pathways”

For the initial analyses I am going to use the following parameters:

'Hugo\_Symbol', 'Variant\_Classification', 'Tumor\_Sample\_Barcode', 'Patient ID', 'Sample ID', 'Pathway', 'Diagnosis Age', 'Sex', 'TMB (nonsynonymous)', 'Pan-Glioma DNA Methylation Cluster', 'Neoplasm Histologic Type Name', 'Neoplasm Histologic Grade', 'Overall Survival (Months)', 'Overall Survival Status', 'IDH status']

To validate my ML models, I am going to test them on another dataset from glioma patients from TCGA [4]. Unfortunately there is no way for me to test and account for data leakage if it happens (if there are some patients that are included in both my dataset from cBioportal and dataset from TCGA)

# **4. Metadata**

Metadata for this study is available on cBioportal.org and is also stored on my personal computer. It includes among others the description of the type of cancer, cancer study identifier (glioma\_mskcc\_2019), study description, pmid (PubMed ID) and citation information. Each data table has a corresponding meta data file.

# **5. Data Quality**

This dataset is curated and I can also filter the data according to my selection criteria on cBioportal, and I use samples with no missing values for the features that I need.

When merging different tables, because some patients lack some samples, and the missing patient ID leads to filtering these samples out, I had to decrease the sample size from 1047 to 745.

# **6. Data Flow**

**Data Extraction**: Pull data from cBioPortal using the **Glioma MSKCC 2019** study.

**Preprocessing**: The data are curated, so I did not need to clean anything, but preprocessing and normalization was needed (e.g., encoding categorical features). The proper names of the genes need to be checked for the further analysis. Cleaning entries without fully matching information was needed after merging data from different experiments.

**Feature Engineering**: Focus on mutation burden, specific mutations, mutated pathways and methylation profiles related to clinical outcomes, investigate which features are of most importance for which model.

# **Modeling:**

1. Apply hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading using clinical and molecular data.

# Use same algorythm to predict patients survival.

# Use Cox proportional hazard analysis with features selected using hierarchical voting-based feature selection algorythm for survival analysis.

1. Use Random survival forest model to predict patients survival using clinical and molecular data.

**Evaluation**: Assess the performance of different models.

# **7. Data Model**

At the conceptual level, I aim to define the relationships between molecular profiles of glioma patients, treatment they receive and clinical outcomes.

At the logical level, I specify the attributes and detail how entities relate to each other:

**Entities and Attributes:**

* **Patients**: Patient ID, age, sex, WGO Grade
* **Molecular Pathways**: Patient ID, Sample ID, Hugo\_Symbol, TMB, Pathway
* **Clinical Outcomes**: Overall survival, Overall survival time

**Relationships:**

* **One-to-Many** between **Patients** and **Genes**.
* **One-to-Many** between **Patients** and **Pathways**.
* **One-to-One or One-to-Many** between **Pathways** and **Clinical Outcomes**.

At the **physical level**, there are no specific requirements ans the dataset I am using is relatively small and I don’t need any specific infrastructure.

# **8. Documentation**

I use Jupiter notebook on Google Colab, saved both on the Google Drive and on my computer, with comments on the code so that my study can be reproduced. This project has a dedicated github repository (https://github.com/GalinaGl/CAS\_ADS\_2024/tree/main/Final%20project)

# **9. Risks**

Glioma is not a very frequent cancer type so my main concern is not to have the sample size large enough to create a powerful predictive model.

In order to test if my models are generalizable, I use another dataset from glioma patients on TCGA from The Cancer Genome Atlas (TCGA) Research Network ([**https://www.cancer.gov/tcga**](https://www.cancer.gov/tcga)). Unfortunately, there is no way for me to test and correct for data leakage that might affect validation of the quality of my models.

# **10. Exploratory data analysis**

After matching data frames from different parts of studies I ended up with 745 patients instead of 1047 I had in the first table. Out of 745 patients 440 are males, 305 are females.

One of the first diagnostic features in glioma is IDH status. Depending on the study and the dataset, 30-70% of glioma patients carry mutations in IDH1 or IDH2 gene. In our dataset we can see that a little more than 50% of the patients contain IDH mutations. IDH mutations are associated with better prognosis, and it seems to be represented on the Fig. 2: we see higher fraction of tumor samples that have IDH mutation in Grade 2 and Grade 3 tumors.

Another characteristic feature is tumor mutational burden (TMB). It is an emerging biomarker for the prediction of immunotherapy success for solid tumors. In the same time, it is associated with poor prognosis in glioma. We can see that TMB is significantly higher in more advanced tumors (Fig. 3).

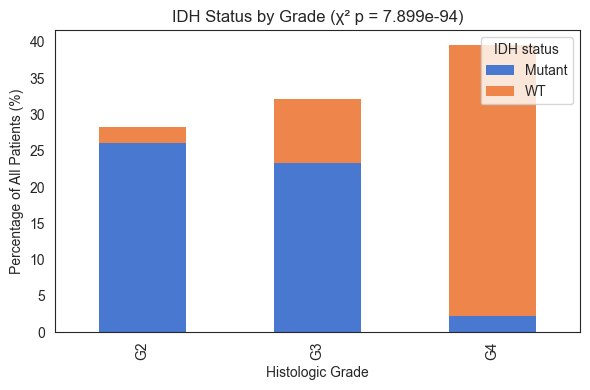


Figure 2: Distribution of IDH mutations by tumor grade (Chi-square test)

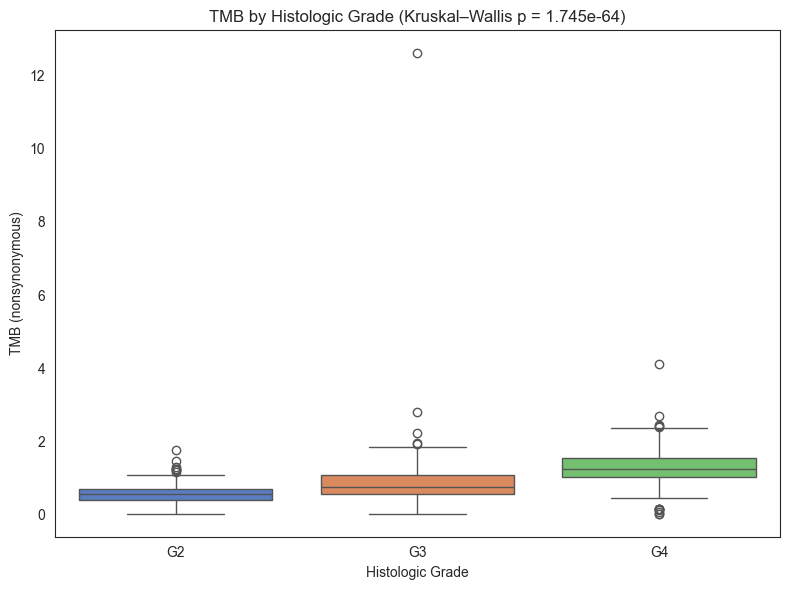


Figure 3: Relationship between tumor mutation burden and WHO Grade Group, Kruskal-Wallis test.

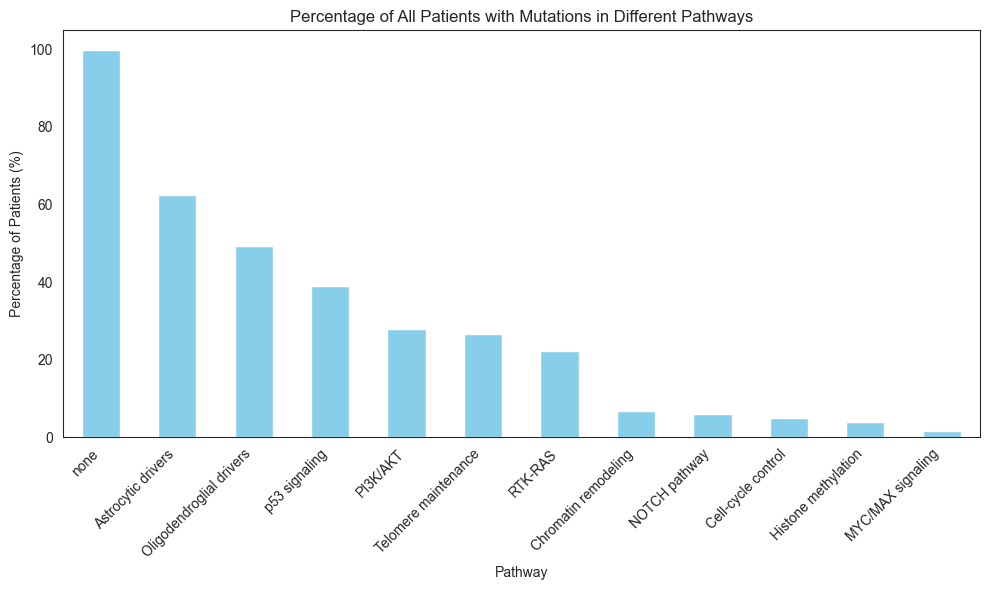


Figure 4: Percentage of mutations in the subset of pre-defined signaling pathways.

Another way to look at the biology underlying certain cancer progression processes is to map the mutated genes to different cellular processes. I mapped the mutated genes to a list of predefined signaling pathways. We can see that the majority of mutations were not assigned a pathway, and the most present pathways were telomere maintenance (which is one of the first stages of carcinogenesis), then there is a number of proliferative signaling pathways that are typical for malignant transformations are targeted (Fig. 4). Most importantly, we can see some differences in the pathways affected depending on WHO grade and diagnosis (Fig. 5), which can teach us something about brain cancer progression.

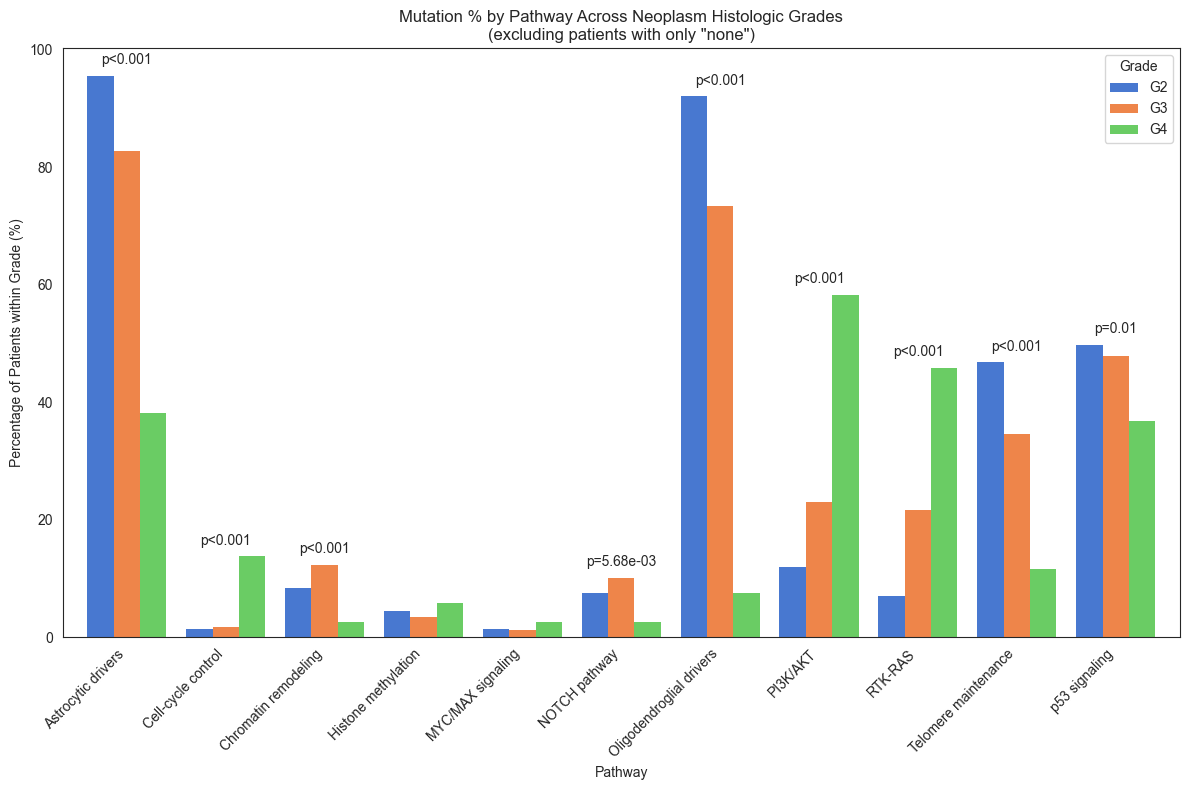


Figure 5: Distribution of mutations in the subset of pre-defined signaling pathways depending on the WHO grade. Chi-2 test.

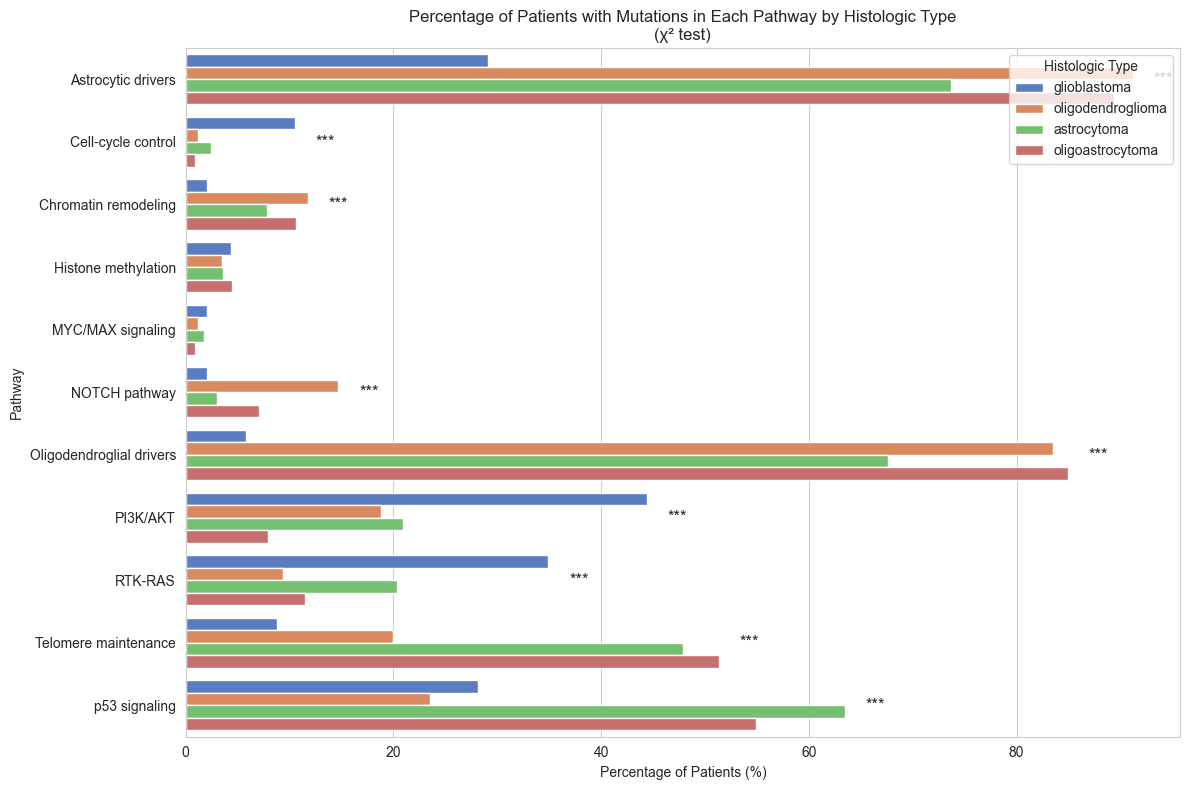


Figure 6: Distribution of mutations in the subset of pre-defined signaling pathways depending on the tumor type. Chi-2 test.

As expected, PI3K/AKT pathway and RTK-RAS pathway that control fundamental hallmarks of cancer: cell proliferation, survival, metabolism and migration have higher mutation rates in G4 tumors. This suggests that dysregulation of one or both of these cascades is a driver of the aggressive biology that defines glioblastoma.

In lower‐grade gliomas, TP53 loss and ATRX‐driven ALT telomere maintenance are early, “founding” events that are nearly universally retained, so their relative frequency is highest in G2/G3 tumors. Grade 4 glioblastomas, however, more commonly activate TERT‐promoter mutations and other proliferative pathways, diluting the prevalence of p53 and ALT‐associated alterations.

NOTCH pathway alterations peak in grade 3 oligodendrogliomas because this subtype relies on Notch signaling to promote the survival and differentiation blockade of mutant oligodendrocyte progenitors. As tumors progress to higher grades or diverge into alternate lineages, other oncogenic cascades (e.g. PI3K/AKT, RTK-RAS) become more dominant, reducing the relative incidence of NOTCH mutations (Fig. 6). Overall the mutation patterns in my dataset are expected for glioma patients [2,3].

I want to also explore the effect that pathways affected have on patients survival. The example I am showing in this report is Kaplan-Meier plot [5,6] comparing overall survival between the patients carrying IDH1 mutations and patients that don’t (Fig. 7). As we saw in Fig. 3, IDH mutations define a distinct, less aggressive glioma subtype with slower growth kinetics and better response to therapy, so the long-term outcomes are much better for those patients. The non-overlapping 95 % confidence bands soon after diagnosis indicate that this difference is highly statistically significant. Other survival plots by pathways affected can be found in my notebook [7].

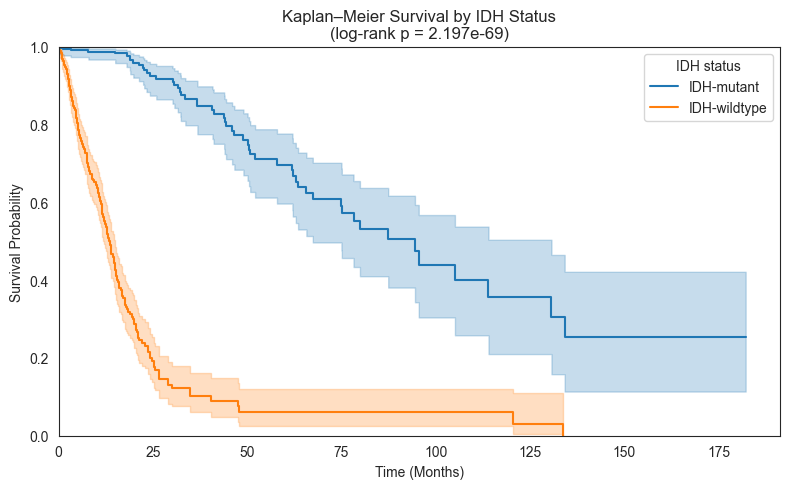


Figure 7: Kaplan–Meier curves of overall survival in glioma patients stratified by IDH mutation status. Shaded bands denote the 95 % confidence intervals around each curve. A log-rank test confirmed a highly significant survival advantage for the IDH-mutant group.

To analyse what clinical and molecular characteristics contribute to survival, I run Cox proportional-hazards model (Fig. 8). Wild-type IDH status and higher histologic grades (G3, G4 compared to G2) are the strongest adverse predictors (CIs do not cross 1). Mutations in the telomere-maintenance and p53-signaling pathways also remain independently prognostic. Mutations in chromatin remodeling genes have a trend toward more favorable survival, whereas PI3K/AKT, RTK-RAS, cell-cycle, NOTCH, and glial-driver pathway alterations, as well as diagnosis age, show CIs overlapping 1, indicating no significant effect in the multivariable model.

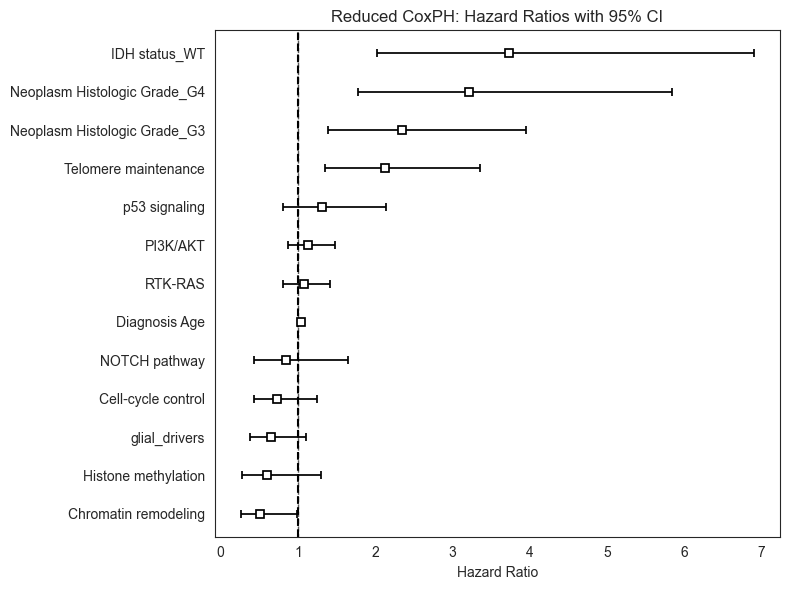


Figure 8. Forest plot of adjusted hazard ratios (squares) with 95 % confidence intervals (horizontal lines) from the reduced multivariable Cox proportional-hazards model for overall survival. Concordance index 0.84.

Then I look at the 20 most frequently mutated genes that I want to use as features for machine learning (Fig. 9). IDH1 sits at the top (∼50 % of patients), confirming its role as the canonical “founder” mutation in lower-grade and secondary gliomas. TP53 (∼38 %) and ATRX (∼27 %) follow closely, reflecting early loss of genome-stability checkpoints and activation of the ALT telomere-maintenance mechanism. PTEN (∼13 %) and EGFR (∼12 %) are the most common true drivers of proliferative signaling outside of IDH-mutant tumors, underscoring PI3K/AKT and RTK/RAS pathway dysregulation in disease progression. Mutations in very large genes **(TTN, MUC16)** are abundant but likely represent passenger burden rather than actionable targets, whereas mid-frequency hits in **PIK3CA**, **NF1**, and **PDGFRA** point to smaller subgroups who might benefit from targeted inhibitors.

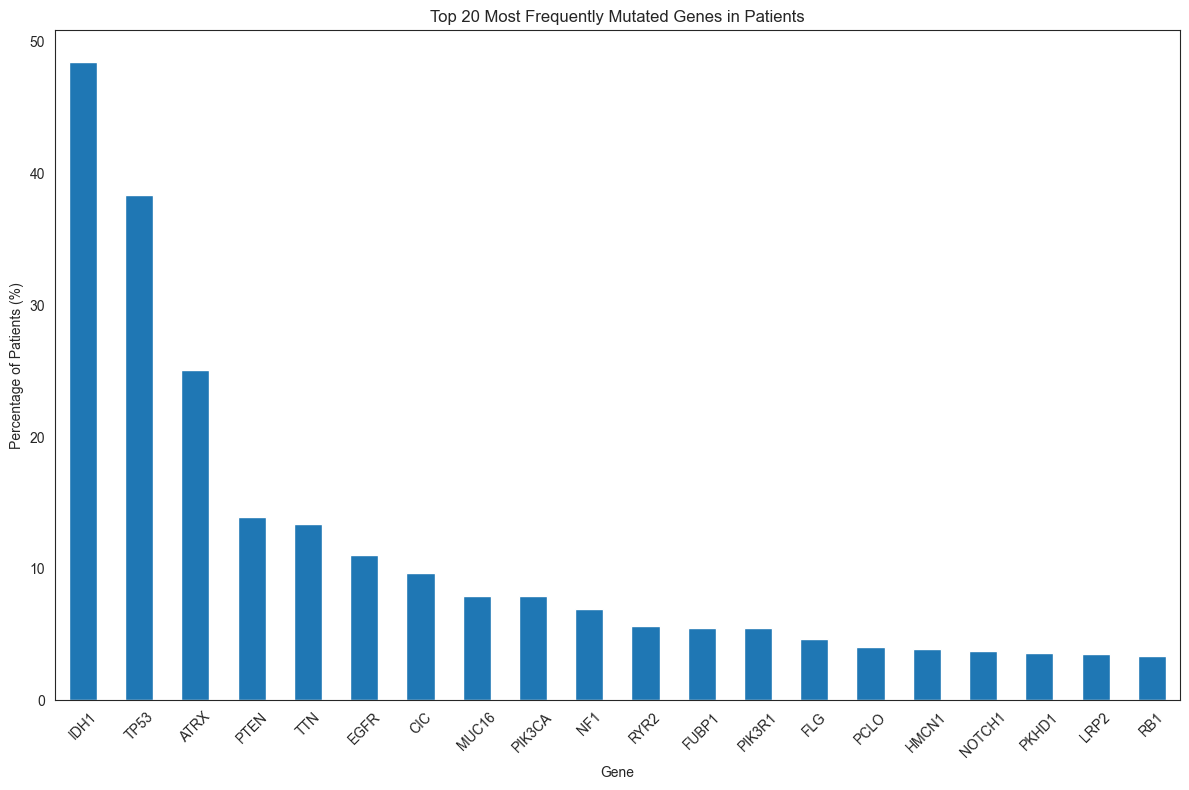


Figure 9: Top 20 most frequently mutated genes in the dataset.

**11 Machine learning analysis**

**11.1 Clustering**

In order to discover molecular subtypes of the brain cancer based on mutation patterns and clinical features I use clustering. As features I use demographic features (age at diagnosis, sex), outcome variables (survival status, survival months), clinical features (tumor type and grade), and molecular features (20 top mutated genes, methylation groups, mutated pathways).

**11.1.1. Hierarchical clustering**

I want to start with hierarchical clustering, and for dimentionality reduction prior to clustering I want to apply PCA. Based on cumulative explained variance I use 12 principal components (90% of variance explained). After inspecting the dendrogram I chose distance cut-off 12 that gives 6 well defined clusters from 52 to 153 patients in each (Fig. 10).

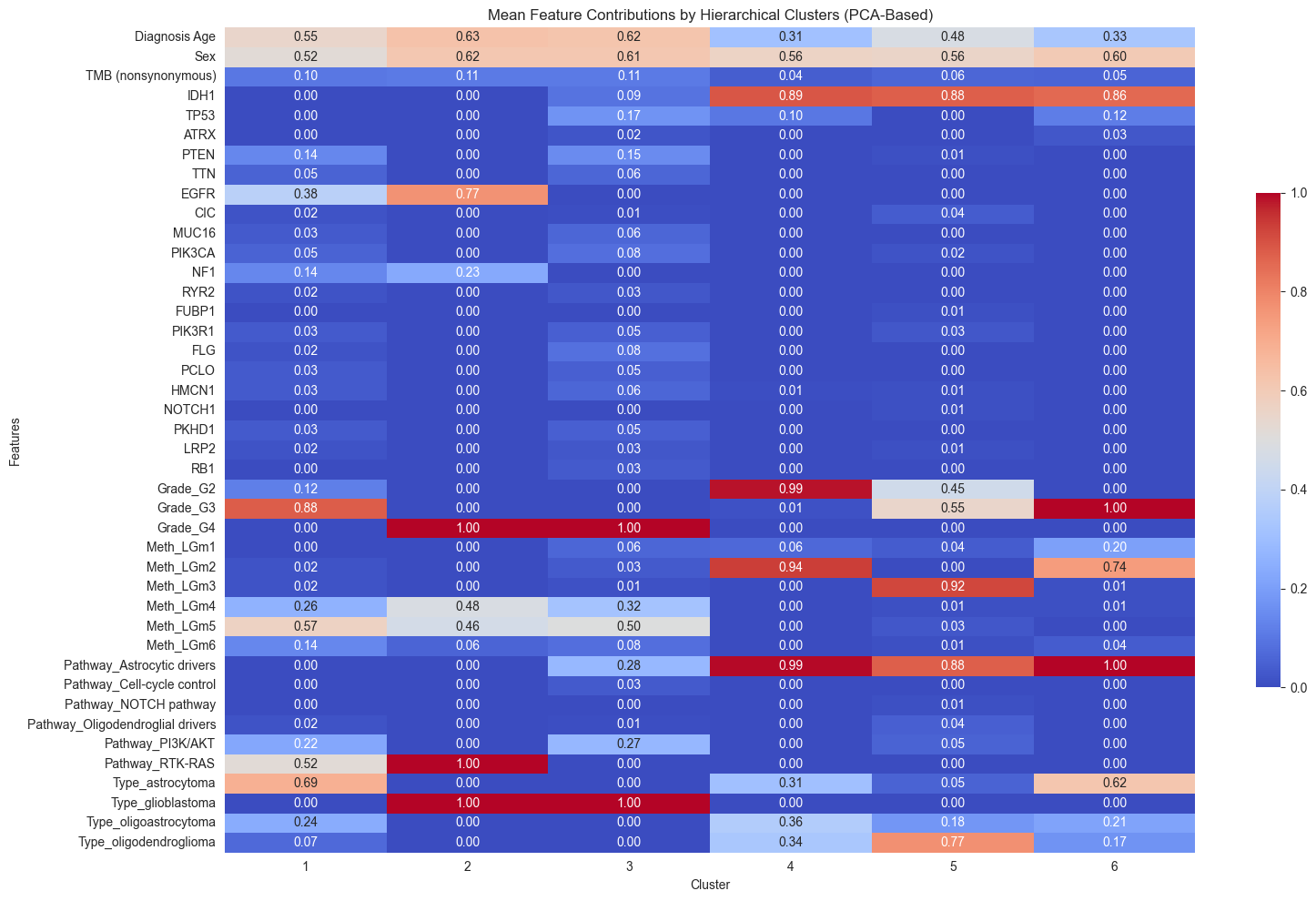


Figure 10. Heatmap of z-score–normalized mean feature values by hierarchical cluster (columns 1–6). Columns 1–6 correspond to the six identified patient clusters, with red indicating above-average feature values and blue below-average.

We see **Cluster 1 (IDH-WT, Grade 3 astrocytomas, moderate age), d**riven by LGm5 methylation (0.57), **RTK-RAS** (0.52)/**PI3K/AKT** (0.22) signaling, with frequent **EGFR** (0.38), **NF1** (0.14) and **PTEN** (0.14) alterations. **Cluster 2 (Classic IDH-WT Grade 4 glioblastoma):** strong **EGFR** (0.77) and **RTK-RAS** (1.00) activation, elevated LGm4 (0.48)/LGm5 (0.46) methylation. C**luster 3 (EGFR-low, astrocytic-driver, IDH-WT Grade 4 glioblastoma):** **PTEN** (0.15)/**PIK3CA** (0.08) hits and substantial LGm4/LGm5 methylation (0.32/0.50). **Cluster 4 (IDH-mutant, Grade 2 astrocytomas, young age): u**niversally high **astrocytic‐driver** score (0.99) and increased ALT **telomere‐maintenance, e**nriched for LGm2 methylation (0.94) and mixed astro/oligo histology. C**luster 5 (IDH-mutant, Grade 2/3 oligodendrogliomas, mid-age): s**trong **oligodendroglial** lineage (0.77), LGm3 methylation (0.92). **Cluster 6 (IDH-mutant, Grade 3 astrocytomas, younger age):** highest **astrocytic‐driver** score (1.00) and LGm2 methylation (0.74), predominantly astrocytoma histology (0.62). More detailed feature contributions to clustering can be found in the notebook.

**11.1.2. Non-hierarchical clustering after UMAP**

When I use non-hierarchical clustering on the data after UMAP dimentionality reduction, I sadly see that DBSCAN gives me 46 clusters with Silhouette Score = 0.60366290807724 and Davies-Bouldin Index = 1.5266057491382032, which shows poorly defined clusters. Then I tried to define which number of clusters from 6 to 16 has the highest Silhouette score and DBI, and 16 clusters give the best results (Fig. 11)

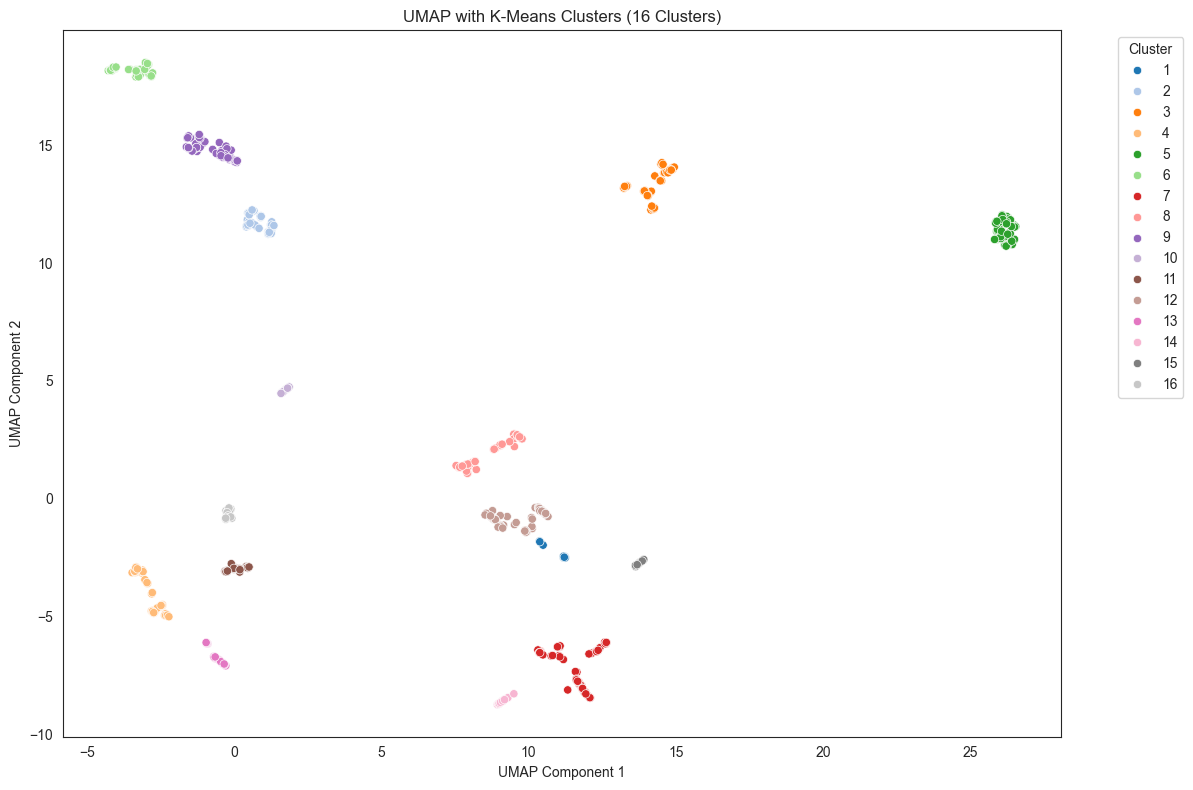


Figure 11. MAP dimensionality reduction of patient feature profiles, with points colored by their assigned K-means cluster (1–16). Each dot represents one patient, positioned by the first two UMAP components. Clusters form discrete groupings in low-dimensional space, reflecting underlying molecular and clinical similarity.

**When we look at the feature contribution heatmap (see notebook), we can see IDH-wildtype, EGFR/RTK-driven GBM clusters (2,3,5,9)**, **IDH-mutant astrocytoma clusters (1, 4, 6, 8, 10, 11, 13, 15, 16** all have IDH1≈0.8–1.0 and Grade\_G2/G3, feature high Pathway\_Astrocytic\_drivers, moderate TP53/ATRX in some, and younger age (Diagnosis Age 0.31–0.50 vs. 0.65+ in GBM clusters), **Oligodendroglioma clusters (7, 12**: Type\_oligodendroglioma≈0.8–1.0, IDH1≥0.8, Grade\_G2/G3, CIC/FUBP1 hits and LGm3/LGm2 methylation subtypes), m**ethylation-dominant subgroups (Cluster 14** stands out with LGm2=1.0, LGm3=0.95, but low pathway and driver‐gene scores; and **Cluster 16** has mixed astrocytic drivers=1.0 and LGm2=0.95 but moderate grade and IDH1=0.95).

Across all 16 clusters we can see the classical triad (IDH-wt GBM vs. IDH-mut astrocytoma vs. oligodendroglioma) broken into finer subgroups by driver pathways (EGFR/RTK vs. astrocytic), methylation subtype (LGm1–LGm6), and histology.

**11.2 Diagnostics**

**11.2.1 Tumor grading**

Tumor grading is a critical step toward treatment optimisation to increase survival rate. I am going to use the same molecular markers as in the previous part to create machine learning model for glioma grading [5,6]. I am going to use a published [4] hierarchical voting-based methodology for feature selection and soft-voting-based ensemble learning model .

This pipeline builds an ensemble feature‐selection process by applying four complementary methods: Weight of evidence scoring, recursive feature elimination with logistic regression, random‐forest importances, and LASSO coefficients, each using class‐balanced sample weights. It then retains only those predictors that at least three out of four methods agree are informative, yielding a conservative, high‐confidence feature set for the next step.

Then it trains five different classifiers (logistic regression, random forest, SVM, k-nearest neighbors, and AdaBoost) using class-balanced sample weights, and combines them into a soft-voting ensemble to leverage their complementary strengths. It evaluates each model’s performance via 10-fold stratified cross-validation, reporting mean accuracy and standard deviation, and then assesses the voting ensemble on the combined holdout predictions with metrics including accuracy, F1-score, confusion matrix, and AUC-ROC.

The final result is summarised on the Figure 12.

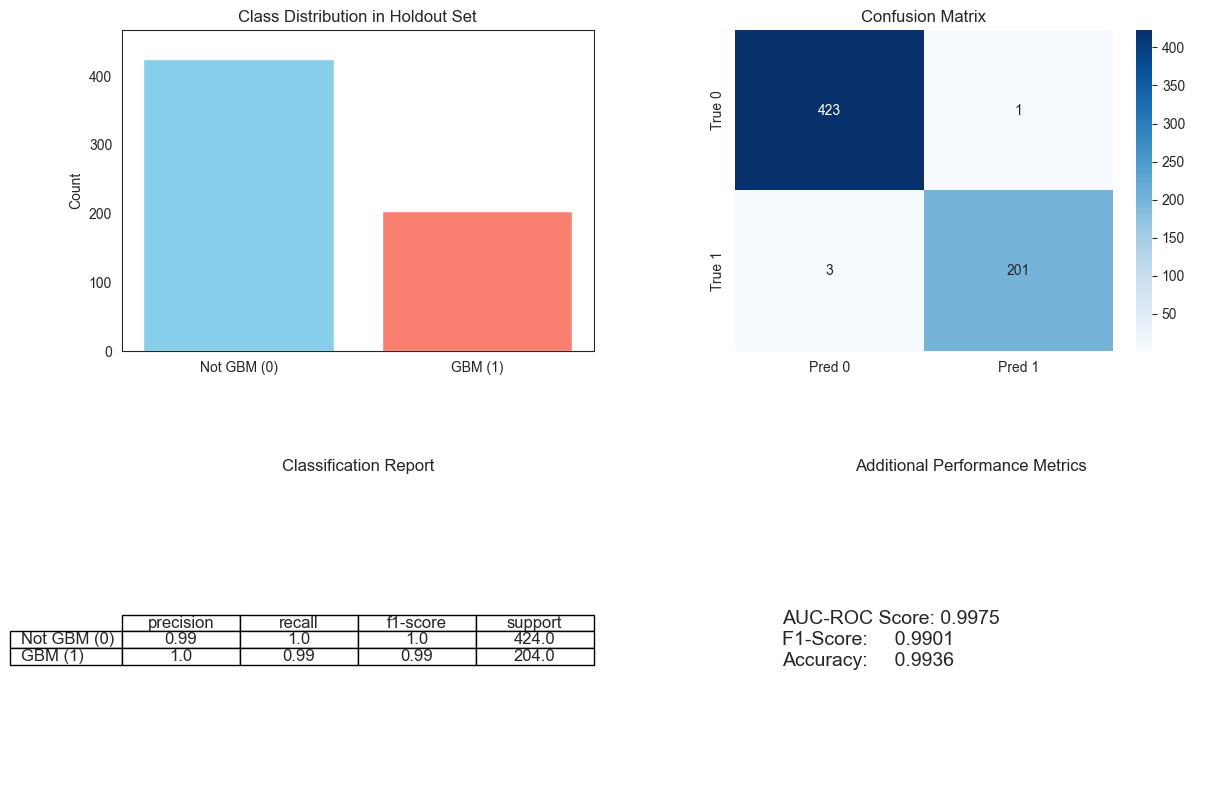


Figure 12. Composite performance summary for the holdout evaluation of the voting ensemble. **Top-left:** class distribution. **Top-right:** confusion matrix. **Bottom-left:** classification report with precision, recall, F1-score, and support for each class. **Bottom-right:** key scalar metrics.

In cross‐validation, the voting ensemble distinguished GBM from non‐GBM almost flawlessly (Accuracy ≈ 0.994, AUC ≈ 0.998), demonstrating that our selected feature set and model architecture have strong predictive power on the development cohort. However, true generalizability remains to be confirmed once the fully independent test set is evaluated.

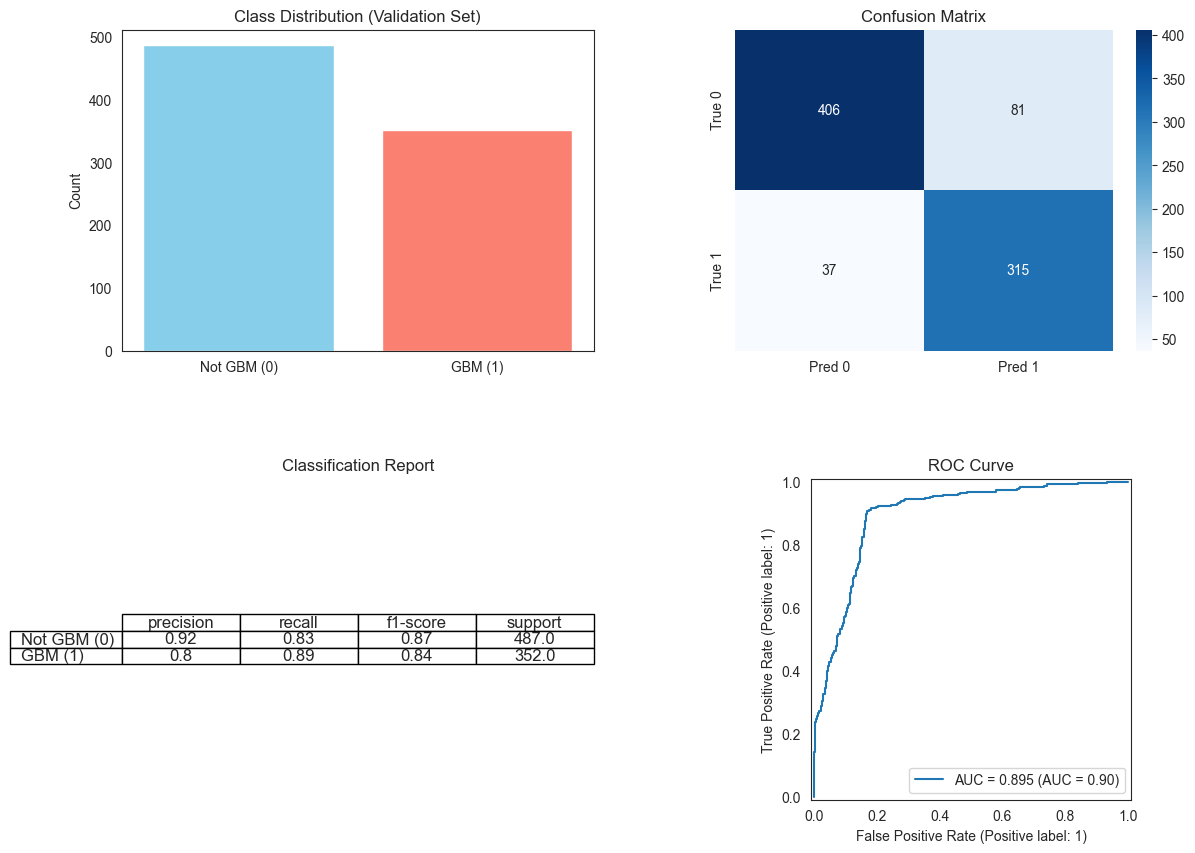


Figure 13. Validation performance of the voting ensemble on the independent TCGA set. **Top‐left:** bar chart showing class distribution. **Top-right:** confusion matrix. **Bottom-left:** classification report table. **Bottom-right:** AUC ≈ 0.895, summarizing the model’s discriminative ability in the validation cohort.

The model’s performance on the independent TCGA set (AUC ≈ 0.90, accuracy ≈ 0.88) confirms that it works reasonably well (Fig. 13), but the drop from cross‐validation scores indicates residual overfitting. Optimising number of features and training on larger datasets might be needed for better generalization.

**11.3 Survival**

Last part of my project is to use the same features to predict survival. For that I first used the same model I used for diagnostics (see notebook), but the performance of the model was much worse than in diagnostics (Table 1).

|  |  |
| --- | --- |
|  | CV Accuracy |
| Random Forest | 0.70 ± 0.13 |
| SVM | 0.62 ± 0.14 |
| KNN | 0.70 ± 0.12 |
| Gradient Boosting | 0.71 ± 0.16 |
| Logistic Regression | 0.64 ± 0.15 |
| Ensemble Voting Classifier (5 models) | 0.74 ± 0.16 |
| Ensemble Voting Classifier (KNN,SVM, GB) | 0.73 ± 0.16 |

I then tried to use Random Survival forest on the entire features set, onon my original dataset split into training and test. Concordance Index on Test Set: was 0.8262, which is considered a good predictive ability, and is better than any setting of my ensemble model. But in order to improve it I investigated feature importance (Fig. 14) and removed features that don’t contribute to the power of the model. Unfortunately this did not change the Concordance Index on Test Set that remained 0.8262.

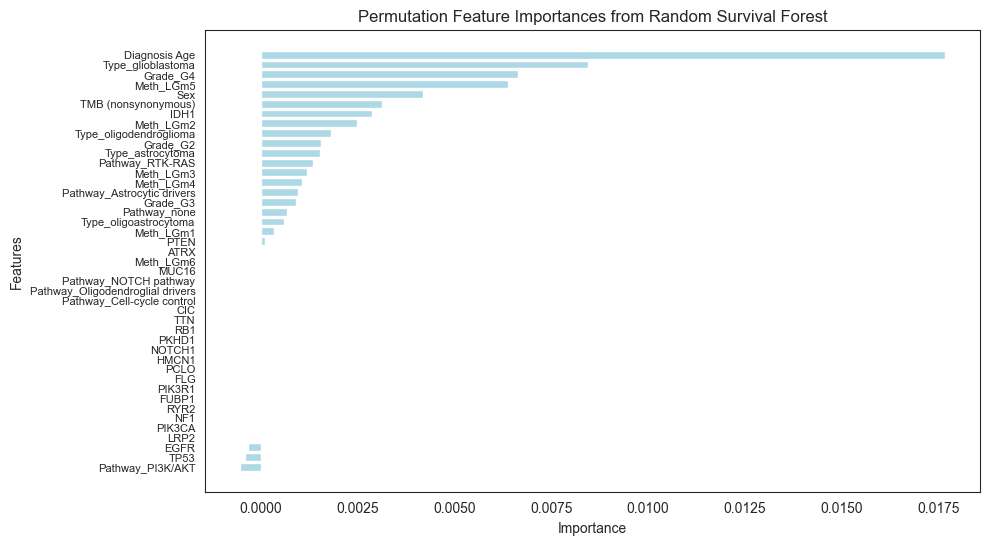


Figure14. Permutation‐based feature importances derived from a random survival forest. Each horizontal bar represents the average decrease in model performance (importance) when that feature is randomly shuffled: longer bars indicate greater impact on survival prediction. Features are sorted in descending order of importance.

# **12. Discussion**

This project demonstrates the power and limitations of integrating molecular profiling into both unsupervised and supervised analyses of adult gliomas. First, I confirmed known biological patterns with descriptive statistics: roughly half of tumors harbored IDH1/2 mutations, which were far more prevalent in lower‐grade (G2/G3) gliomas and correlated with significantly better survival. Meanwhile, tumor mutational burden (TMB) increased with histologic grade, underscoring its role as a biomarker of genomic instability and aggressive behavior. Pathway‐level mapping revealed that glioma initiation and maintenance are driven more by cell‐lineage programs (astrocytic and oligodendroglial driver mutations) than by downstream proliferation pathways. The intermediate prevalence of p53, PI3K/AKT, and telomere‐maintenance alterations suggests they act as cooperators rather than primary instigators.

The unsupervised clustering analyses (both hierarchical after PCA reduction and K-means after UMAP reduction) uncovered molecularly and clinically coherent subtypes beyond the classic triad of IDH-wildtype glioblastoma, IDH-mutant astrocytoma, and IDH-mutant oligodendroglioma. The six clusters from PCA‐based hierarchical clustering neatly partitioned by IDH status, grade, histology, and driver pathways (e.g. high EGFR/RTK activation in classic GBM vs. ALT telomere maintenance and astrocytic drivers in IDH-mutant tumors). Expanding to 16 K-means clusters further subdivided these groups according to methylation subtypes (LGm1–LGm6) and additional pathway activity, suggesting the existence of fine‐grained molecular niches that may respond differently to targeted therapies. These findings align with the 2016 WHO move toward integrated molecular diagnostics and hint at even deeper substructure that could inform precision treatment [2,3].

Survival analyses reinforced the prognostic importance of key molecular events. Kaplan–Meier curves showed a highly significant survival advantage for IDH-mutant patients, with non-overlapping 95 % confidence bands. The multivariable Cox model pinpointed wild-type IDH and higher WHO grade (G3/G4) as the strongest adverse predictors, while telomere‐maintenance, p53‐signaling and chromatin remodeling mutations remained independently prognostic. Other pathways (e.g. PI3K/AKT, RTK-RAS) and clinical variables (age, sex) showed confidence intervals crossing unity, indicating their effects may be mediated by or correlated with the dominant drivers.

In the supervised learning, the hierarchical voting‐based feature selection produced a conservative “core” feature set drawn from weight-of-evidence, recursive feature elimination, random‐forest importance, and LASSO. With the soft‐voting ensemble of five classifiers (logistic regression, random forest, SVM, KNN, AdaBoost) I achieved very good discrimination in cross‐validation (accuracy ≈ 0.994, AUC ≈ 0.998), underscoring the predictive richness of the selected features on the development cohort. However, validation on an independent TCGA set yielded a reduced but respectable AUC ≈ 0.895 and accuracy ≈ 0.88, with precision/recall declines particularly in the non-GBM class (0.92/0.83). This drop indicates residual overfitting and suggests that further regularization, simplified model architectures, or expanded training data may be needed for robust generalization.

Finally, for survival prediction, a random survival forest achieved a concordance index of 0.826 on a held-out test set (better than the diagnostic ensemble’s survival classification) validating its utility for modeling time‐to‐event outcomes. Permutation importance again highlighted age, histologic grade, and key pathway mutations as the top prognostic features, whereas many individual gene alterations (e.g. large passenger genes such as TTN, MUC16) contributed negligibly [5,6].

**13. Conclusions**

Incorporating molecular markers (particularly IDH status, telomere‐maintenance, and p53 pathway alteration) substantially refines glioma classification beyond histology alone, yielding clinically and biologically coherent subtypes with distinct prognoses. Unsupervised clustering reveals further subtypes stratified by driver pathways and methylation profiles, raising the possibility of tailoring therapies to these finer molecular niches. A conservative voting‐based feature selection and ensemble classifier achieves great performance on the development data but experiences performance decay on independent validation, highlighting the need for additional data, cross‐study harmonization, and potentially simpler models to ensure real‐world generalizability. Random survival forests leveraging the same core features demonstrate good concordance in predicting patient outcomes, supporting their role in prognostic modeling and risk stratification.

To close the generalization gap, the model should be trained on larger, multi‐institutional cohorts, ideally with prospective validation, and benchmarked against clinical risk scores. Integration of additional data modalities (e.g. imaging, transcriptomics), dynamic monitoring (e.g. serial liquid biopsy), and exploration of treatment‐response phenotypes will further advance precision medicine in glioma care.

In summary, this project confirms that molecular profiling is indispensable for modern glioma diagnosis and prognosis, and that thoughtful feature selection combined with ensemble learning can yield powerful predictive tools, provided they are rigorously validated and continuously refined with broader, real‐world data.

Bottom of Form

# Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 07.07.25 Signature(s):



# References and Bibliography

[1] Genomic Correlates of Disease Progression and Treatment Response in Prospectively Characterized Gliomas. Clin Cancer Res. 2019 Sep 15;25(18):5537-5547. doi: 10.1158/1078-0432.CCR-19-0032.

[2] The 2021 WHO Classification of Tumors of the Central Nervous System. Neuro-Oncology 23(8), 1231–1251, 2021 | doi:10.1093/neuonc/noab106

[3] Molecular markers in Gliomas: A practical review and algorithm proposal, Interdisciplinary Neurosurgery, Volume 41, 2025, 102062, ISSN 2214-7519. https://doi.org/10.1016/j.inat.2025.102062.

[4] Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics. Int. J. Mol. Sci. 2022, 23, 14155. https://doi.org/10.3390/ijms232214155

[5] Uses of artificial intelligence in glioma: A systematic review. Medicine International 4, no. 4 (2024): 40. https://doi.org/10.3892/mi.2024.164

[6] Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC Med Res Methodol* **23**, 268 (2023). <https://doi.org/10.1186/s12874-023-02078-1>

[7]https://github.com/GalinaGl/CAS\_ADS\_2024/blob/main/Final%20project/Data/Final\_project\_GGlousker.ipynb