International Journal Of Public Mental Health And Neurosciences
ISSN: 2394-4668
(An Official publication of Sarvasumana Association)

Machine Learning Driven Subtype and Mutation Classification in Gliomas Using Transcriptomic Data

Shreeya Pahune and Bhaswar Ghosh
Center for Computational Natural Sciences and Bioinformatics
Institute of Information Technology - Hyderabad
Hyderabad, India
shreeya.pahune@research.iiit.ac.in; bhaswar.ghosh@iiit.ac.in

***Abstract*— Gliomas, accounting for nearly one-third of all brain tumours, present themselves as significant medical challenges owing to their heterogeneous and aggressive nature. Given that traditional diagnostic techniques often fall short of capturing the entire genetic landscape, there are minimal options currently present in personalised medicine, thus adding to existing challenges. In order to aid precision oncology, we present a machine learning (ML) framework for glioma subtype classification and mutation prediction using RNA sequencing data. We used a two-step feature selection procedure that included XGBoost and LASSO, leveraging RNA sequencing and somatic mutation data of 667 glioma samples from the Cancer Genome Atlas (TCGA) Pan-Cancer project to identify key genes that are associated with glioma subtypes and important mutations. Metrics such as AUROC, F1-Score, and Matthews Correlation Coefficient (MCC) were used to train and evaluate the supervised machine learning models, including Random Forest and Stochastic Gradient Descent. Our framework performed robustly despite the class imbalance present in the dataset. We achieved a high classification accuracy for Diffuse Glioma (DG) subtypes (AUROC 0.90-1.00), distinguishing between low-grade gliomas (LGGs) and glioblastoma multiforme (GBM). Mutation status predictions for the key prognostic genes, including IDH, TP53, and ATRX, achieved high AUROC scores of 0.89-0.98 and MCC values of 0.72-0.97. Further, we used feature sets from our top-performing models to perform gene set enrichment as part of our post-analysis. This confirmed the biological significance of identified genes related to established carcinogenic pathways and further validated the clinical applicability of our technique. Our ML-driven framework presented offers a scalable, data-driven solution in computational oncology with potential applications across cancer types.**

***Index Terms*—Glioma Classification, Mutation Prediction, Machine Learning, RNA Sequencing, Precision Oncology**

I. Introduction

Glial cells, which are crucial for maintaining neurons and their surroundings, are the primary cause of gliomas, the most common tumours in the brain and spinal cord [1]. As a standard, the World Health Organization (WHO) assigns a grade of I, II, III, or IV to tumours. However, regardless of their grade, gliomas are known to grow and cause disability or even death [2]. Generally, adult diffuse gliomas (DGs) comprise gliomas of grades II, III, and IV and can be further categorised into low-grade gliomas (LGGs), which include grades II and III, and glioblastoma multiforme (GBM), grade IV, the most invasive and deadly glioma [3, 4].

Histopathology has historically been used to diagnose and classify gliomas. This classification was well-established and served as the foundation for the WHO classification of central nervous system tumours, which has evolved significantly through multiple editions, with the latest update published in 2021 [2]. However, it has significant intra- and inter-observer variation, especially for grade II–III tumours (LGGs) [5]. The classification of gliomas into subtypes based on genomic, epigenomic, and proteomic profiles has come a long way as well. Discoveries of multiple biologically and prognostically important biomarkers have led to new classifications of gliomas. Among the several key mutations involved in gliomas, isoforms of isocitrate dehydrogenase (IDH1 and IDH2) are very common [6]. Tumour protein p53 (TP53), a known glioma driver [7], is another such mutation that is often dysregulated in cancer. Alpha-thalassemia/mental retardation, X-linked (ATRX) is a chromatin-remodelling protein that is encoded by the X chromosome and was recently discovered as a clinical target [8]. Mutations in the ATRX genes are a common occurrence in LGGs. The development of advanced sequencing and data integration methods has made molecular profiling an attractive tool for detecting patterns and markers unique to tumours. Our proposed method is more reliable than traditional histopathological evaluations as it has the potential to improve diagnostic accuracy in glioma subtyping and grading by identifying prognostic markers.

Based on histological classification, LGGs were formerly divided into Astrocytoma (A), Oligoastrocytoma (OA), and Oligodendroglioma (OD). However, they exhibit extremely diverse clinical behaviour, which makes it difficult to predict subtypes based on histologic class. Recent studies suggest, irrespective of grade or histology, glioma cases should be divided into three categories: IDH wildtype cases, IDH mutant samples containing codeletion of chromosome arms 1p and 19q (IDH mutant-codel), and samples with euploid 1p/19q (IDH-mutant-non-codel) [9, 10]. This classification of LGGs based on histopathology and the one with molecular signature share

© IJPMN, Volume 12, Issue 2, August-2025

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

similarities; for example, LGGs with both an IDH mutation (i.e., a mutation in either IDH1 or IDH2) and 1p/19q codeletion occur most often in oligodendrogliomas [10]. Understanding how these mutations affect the pathophysiology of gliomas may help us develop precision treatments that target the pathways most important to the survival and proliferation of each subtype. Based on gene expression and genomic clustering, four gene expression subtypes were found in the 2010 TCGA categorisation of GBM [11], namely, Classic (C), Mesenchymal (M), Proneural (P), and Neural (N). Additionally, nearly a third of GBM patients harbour a specific deletion known as epidermal growth factor receptor variant III (EGFRvIII), an independent poor prognostic predictor. In contrast, 50% to 60% of patients have overexpressed EGFR [12]. This goes to show that molecular analyses have further helped advance our understanding of the biology behind gliomas and transformed the way we diagnose them. There may be unique genetic and clinical downstream effects when a particular mutation is present, and these advancements have the potential to capture them and thus identify new treatments and strategies that target specific glioma subtypes [13].

In this study, we present a complete framework for a thorough understanding of gliomas, including its subtypes and key mutations (Figure 1). The aim of the machine learning pipeline is to accurately predict the subtypes and mutation status of the glioma-causing genes. To ensure that we can identify the biological underpinnings of these gliomas, we focused on a comprehensive analysis of the mutation and clinical data before the ML workflow as part of our pre-analysis. Additionally, a post-analysis was carried out using the extracted features to understand enriched pathways and processes. Further, the end-to-end pipeline was developed to aid a pathologist in the diagnosis and treatment of gliomas, which can be scaled to other diseases as well.

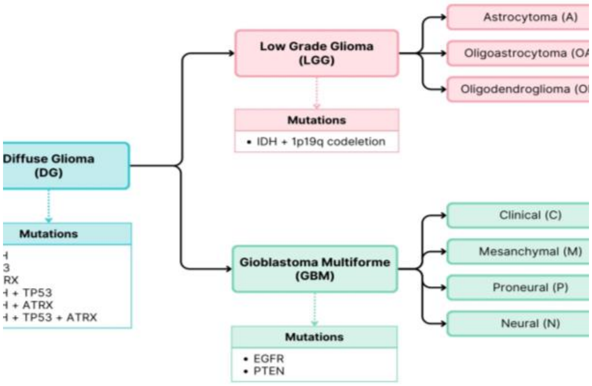


Fig. 1. Overview of glioma subtypes with associated key genes.

II. Results

A. Preliminary Analysis

1) Mutational Landscapes in Glioma Subtypes

Our preliminary analysis aims to investigate the implications of mutations and subtypes within glioma. Figure 2, 3 and 4 presents the mutational landscape of DG, which reveals that IDH1, TP53, and ATRX are among the most frequently mutated genes, with mutation rates of 46%, 39%, and 23%, respectively. Co-occurrence and exclusivity patterns are also evident, indicating that IDH1 mutations co-occur significantly with TP53 and ATRX mutations ($p < 0.05$), while IDH1 shows strong mutual exclusivity with EGFR and PTEN ($p < 0.05$).

For LGG specifically, similar co-occurrence patterns are observed between IDH1, TP53, and ATRX. Other frequently mutated genes included oncogenes like PIK3CA, IDH2 and NOTCH1, as well as tumour suppressor genes such as FUBP1 (also a known oncogene) and NF1. We also noted co-occurrence relationships between CIC and FUBP1.

Among GBM-specific mutations, PTEN (30%), TP53 (approximately 25%), TTN (25%), and EGFR (24%) are the most altered genes, showing significant co-occurrence ($p < 0.05$). IDH1, along with EGFR and PTEN each, exhibits strong mutual exclusivity (p-value < 0.05), while co-occurrence is observed between several genes, such as PI, AHNAK, and PCLO ($p < 0.05$).

C>T transitions are the most common type of mutation, constituting nearly 60% of all DG mutations. T>C changes follow, as seen in the Ti-Tv plot. We also see comparable trends in GBM and LGG. The Ti/Tv ratio reveals that transitions are much greater than transversions. Supplementary file 1 provides the complete mutational landscape for each glioma subtype.

2) Survival Impact

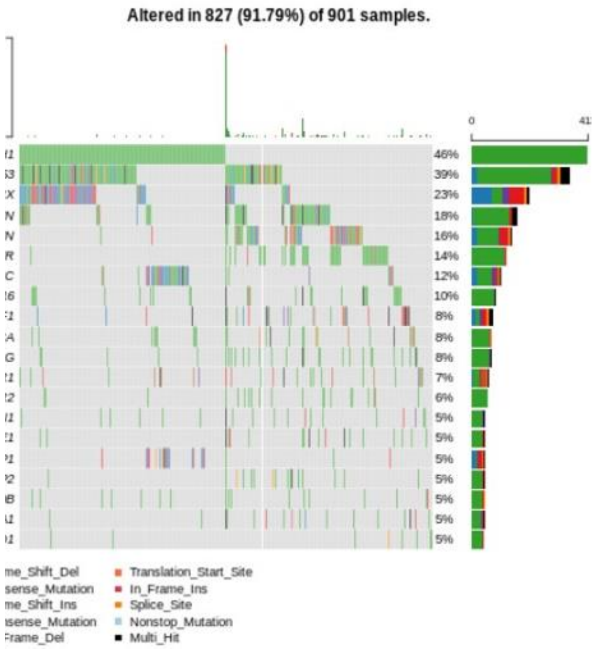
Each subtype and the significant mutations within them were then investigated using clinical data to determine whether they had an advantageous or detrimental effect on the survival of an individual with glioma. Figure 5 illustrates, for DG and LGG, respectively, the survival probability within subtypes (Refer to supplementary file 2 for a complete analysis).

In diffuse gliomas, the co-occurrence of mutations showed striking survival patterns. IDH-mutants with ATRX co-occurrence demonstrated significantly better survival than cases without co-occurrence ($p<0.0001$). Similarly, IDH-mutant cases with TP53 co-occurrence exhibited markedly improved survival outcomes compared to non-co-occurring cases ($p<0.0001$). The survival curves for both molecular combinations showed clear separation, with co-occurring mutations maintaining higher survival probabilities over time. We also noted that IDH mutations generally gave a survival advantage, but their influences increased when co-occurring with TP53 or ATRX. This suggests the possibility that these molecular changes could improve patient outcomes together.

On the other hand, GBM demonstrated no significant differences in survival between molecular subtypes ($p = 0.93$) or

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

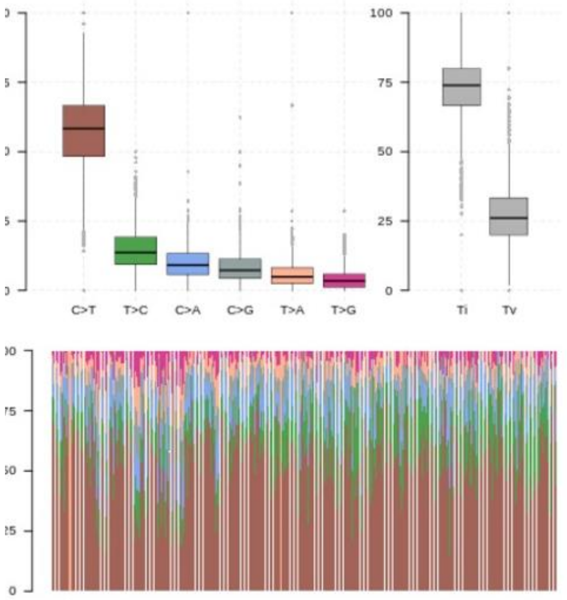
significant mutations such as PTEN (p = 0.74), indicating that



this tumour type (stage IV) is a highly aggressive subtype.

Fig. 2. Mutational Landscape of DG: Onco-plot for the top ten mutated genes.

Fig. 3. Mutational Landscape of DG: Transition and Transversion (Ti-Tv) profile in the mutation landscape



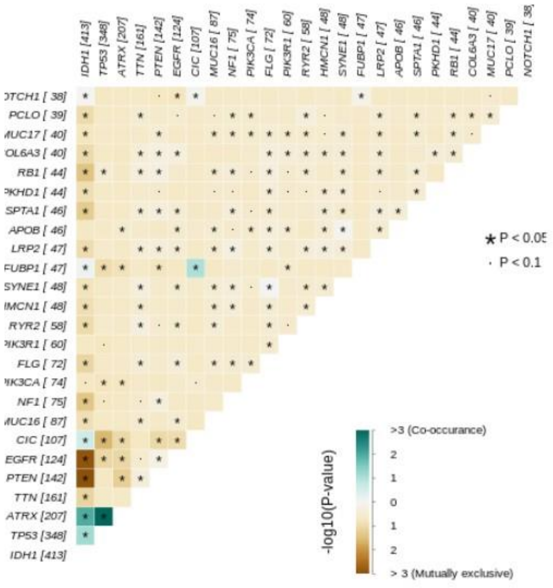


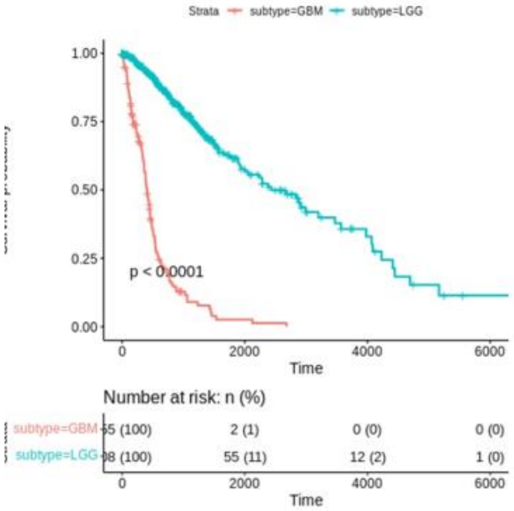
Fig. 4. Mutational Landscape of DG: Pairwise exclusivity or co-occurrence analysis among frequently mutated genes.

The survival time is specified in days till the last day of follow-up or death.

Fig. 5. Survival Analysis: Subtypes of DG: LGG and GBM.

B. Machine Learning Model Performance

Each of our six ML models was trained on each feature set for every subtype and mutation. XGBoost and LASSO methods



(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

were used for feature selection, which allowed us to identify prognostic genes for the majority of models. The union of feature sets from these two approaches was used selectively in two models. A full breakdown of each model's performance with all feature sets across the three glioma types is available in the metrics folder in our GitHub repository. Finally, we leveraged MCC to identify the best model for binary classification, and in other cases, the model with the highest accuracy was considered.

We achieved high performance for DG. As shown in Table 1, the MCC values were between 0.70 and 1.00, and AUROC between 0.90 and 1.00. The Random Forest classifier, along with XGBoost-based feature selection, achieved perfect classification for LGG/GBM subtypes, with AUROC, F1-Score, MCC, Precision, Recall, and Specificity all scoring 1.00. For IDH mutation prediction, the Gaussian Naive Bayes (GNB) classifier using the union of feature sets performed exceptionally well, achieving an AUROC of 0.98 and an F1-Score of 0.99. Predictions when using co-occurring mutations (IDH + TP53, IDH + ATRX, IDH + TP53 + ATRX) also showed good results, with AUROC values ranging from 0.89 to 0.95.

LGG subtype classification (Table 2) into O/OA/OD was moderate, with an accuracy of 65% using the Random Forest classifier with the union of the two feature sets. However, mutation predictions, especially for IDH mutation and 1p/19q codeletion, achieved perfect scores across all metrics, thus emphasising the high reliability in identifying these key biomarkers.

The performance of the models for the GBM subtype is summarised in Table 3. The Nearest shrunken centroids (NSC) classifier using XGBoost-based feature selection achieved 83% accuracy for GBM subtype classification (C/M/N/P) (AUROC: 1.00). However, performance for the mutation prediction task within this subtype was lower due to high-class imbalance, with the AUROC values ranging from 0.72 to 0.76 for EGFR and PTEN mutations.

TABLE I. BEST PERFORMING MODELS FOR DG WITH THE FEATURE SELECTION ALGORITHM

Subtype Classification							
Subtypes	***Classifier***	***Feature Sel.***	***AUROC***	***F1-Score***	***MCC****	***Precision***	***Recall***
LGG/GBM	RF	XGBoost	1	1	1	1	1
Mutation Prediction							
Mutation	***Classifier***	***Feature Sel.***	***AUROC***	***F1-Score***	***MCC****	***Precision***	***Recall***
IDH	G-NB	Union	0.98	0.99	0.97	0.99	0.99
TP53	RF	XGBoost	0.9	0.89	0.81	0.91	0.88
ATRX	LogReg	XGBoost	0.91	0.85	0.79	0.79	0.92
IDH + TP53	LogReg	XGBoost	0.95	0.93	0.9	0.91	0.96
IDH + ATRX	NSC	XGBoost	0.93	0.85	0.8	0.75	0.97
IDH + TP53 + ATRX	NSC	Lasso	0.89	0.77	0.7	0.65	0.93

*represents the metric used for evaluating the models

TABLE II. BEST PERFORMING MODELS FOR LGG WITH THE FEATURE SELECTION ALGORITHM

Subtype Classification							
Subtypes	***Classifier***	***Feature Sel.***	***AUROC***	***Accuracy****	***F1-Score***	***MCC***	***Precision***
O/OA/OD	RF	Union	0.99	0.65	0.7	0.47	0.77
Mutation Prediction							
Mutation	***Classifier***	***Feature Sel.***	***AUROC***	***Accuracy****	***F1-Score***	***MCC***	***Precision***
IDH + 1p/19q codel	RF	Lasso	1	1	1	1	1

*represents the metric used for evaluating the models

TABLE III. BEST PERFORMING MODELS FOR GBM WITH THE FEATURE SELECTION ALGORITHM

Subtype Classification							
Subtypes	***Classifier***	***Feature Sel.***	***AUROC***	***Accuracy****	***F1-Score***	***MCC***	***Precision***
C/M/N/P	NSC	XGBoost	1	0.83	0.83	0.77	0.84
Mutation Prediction							
Mutation	***Classifier***	***Feature Sel.***	***AUROC***	***F1-Score***	***MCC****	***Precision***	***Recall***
EGFR	SGD	Lasso	0.72	0,60	0.42	0.55	0.67
PTEN	SGD	Lasso	0.76	0,67	0.48	0.57	0.8

*represents the metric used for evaluating the models

C. Functional Enrichment Analysis

Supplementary File 3 summarises the main findings of this analysis, which was carried out using Enrichr and DAVID's functional annotation clustering. We identified key pathways and significant gene ontology (GO) terms related to the important mutations and observed a general similarity between the two, thus confirming biological relevance.

IDH mutations in DG are mostly associated with metabolic pathways and activities, such as folate biosynthesis and carbon dioxide transport. TP53 mutations have been linked to immune and cell cycle processes, including death receptor activity and mast cell activation, and they are more common in pathways associated with cancer. TP53 was linked to both the immunological response and the cell cycle. Mutations involving TERT along with IDH and ATRX highlighted the cellular structure and transport functions, especially plasma membrane composition and amine transport. We observed a repetition of vision-related terms and other similar keywords related to IDH, TP53, and ATRX independently were enriched when co-occurring mutations of IDH + TP53, IDH + ATRX, and IDH + TP53 + ATRX were considered. Similar pathways and clusters

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

of annotations, such as phagolysosome and cortical actin cytoskeleton, were highlighted in the cases of PTEN and EGFR.

III. METHODS

A. Data pre-processing

We obtained RNA sequencing (RNA-seq), mutation and clinical data of GBM and LGG from the TCGA Pan-Cancer project [14] using the TCGAbiolinks package [15] in R. For the RNA-seq data, we considered only those primary tumour samples with corresponding somatic mutation data. This resulted in a matrix of 667 samples (LGG = 511 and GBM = 156), each with 56603 genes (Table 4 and 5). The variance-stabilising transformation (VST) was applied across each of the raw RNA-seq matrices to account for the difference in variances contributed by individual genes. Finally, each transformed RNA-seq matrix and mutation data were integrated to create the final dataset for the ML pipeline (Figure 6), where the labels corresponding to each sample's subtype and mutation status were derived from the mutation data. Additionally, we performed preliminary analysis independently using the clinical and mutation data independently for each subtype.

TABLE IV. SAMPLE DISTRIBUTION ACROSS THE SUBTYPES

Diffuse Glioma (n = 667)	
Low-Grade Glioma (LGG)	511
Glioblastoma Multiforme (GBM)	156
Low-Grade Glioma (n = 510)	
Astrocytoma (A)	192
Oligoastrocytoma (OA)	129
Oligodendroglioma (OD)	189
Glioblastoma Multiforme (n = 145)	
Classical (C)	39
Mesenchymal (M)	50
Neural (N)	26
Proneural (P)	30

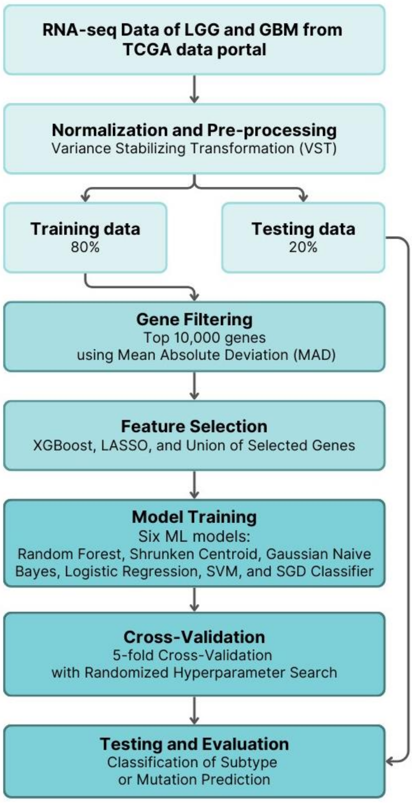
TABLE V. SAMPLE DISTRIBUTION ACROSS THE SUBTYPES

Diffuse Glioma (n = 667)	
IDH [0/1]	[239/415]
TP53 [0/1]	[375/279]
ATRX [0/1]	[475/179]
IDH + TP53 [0/1]	[431/223]
IDH + ATRX [0/1]	[486/168]
IDH + TP53 + ATRX [0/1]	[505/149]
Low-Grade Glioma (n = 508)	
IDHmut + 1p/19q codel	94
IDHmut + non-codel	246
IDH Wildtype (IDHwt)	168
Glioblastoma Multiforme (n = 151)	
EGFR [0/1]	[109/42]
PTEN [0/1]	[103/48]

*0 indicates the absence of the mutation, whereas 1 indicates its presence.

B. Preliminary Analysis

To identify the key mutations for each main glioma, we began by generating a mutational landscape for each subtype using only mutation data. The analysis included creating onco-plots to highlight the top most mutated genes, analysing significant pairwise exclusivity or co-occurrence events, exploring the occurrence and patterns within Ti-Tv mutations, as well as identifying highly affected oncogenic pathways for the top ten mutated genes in each subtype. Following this, we conducted a survival analysis for each subtype, and each identified mutation from the first step was leveraged only from clinical data. This ensured that specific subtypes or the selected mutations had a significant impact, whether advantageous or detrimental to a patient's survival. The Kaplan-Meier method was applied to estimate the survival probability at different time intervals. We considered the length of time from the date of metastatic lung cancer diagnosis to the date of death or last follow-up as the overall survival time. Patients still alive at the time of analysis were censored at the most recent assessment date.


RNA-seq Data of LGG and GBM from TCGA data portal
Normalization and Pre-processing
Variance Stabilizing Transformation (VST)
Training data
80%
Testing data
20%
Gene Filtering
Top 10,000 genes
using Mean Absolute Deviation (MAD)
Feature Selection
XGBoost, LASSO, and Union of Selected Genes
Model Training
Six ML models:
Random Forest, Shrunken Centroid, Gaussian Naive Bayes, Logistic Regression, SVM, and SGD Classifier
Cross-Validation
5-fold Cross-Validation
with Randomized Hyperparameter Search
Testing and Evaluation
Classification of Subtype
or Mutation Prediction

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

Fig. 6. ML pipeline for feature extraction and model building to classify cancer subtypes and predict mutations.

C. ML Pipeline

Our pipeline began by performing a stratified train and test split on the VST transformed matrix by holding out 20% of samples for testing purposes, and the remaining were used for training. Within the training set, we filtered the top 10,000 most variable genes selected using mean absolute deviation (MAD) from the initial 56603 genes. This resulted in our input matrix having the dimensions (n,10000), where n is the number of samples in the glioma under consideration. We extracted prognostic genes from these 10,000 nominated genes by applying the XGBoost and LASSO feature selection methods. In both methods, the genes with feature importance less than 0 were eliminated, and the ranked lists were saved for further post-analysis.

The XGBoost feature selection approach was taken into consideration since we had a large number of genes, and this number was significantly greater than the number of samples. The motivation for using LASSO regression as the alternate technique was eliminating irrelevant genes from a large set of 10,000. This method shrinks coefficients to zero, therefore making it easier to eliminate features that do not contribute to the output. Additionally, to combine complementary features, a union of the two feature sets from the XGBoost and LASSO methods was also used. This was done in cases when the two feature selection algorithms, individually, could not extract features that predicted labels with reasonable accuracy.

Using these feature sets, we then trained six machine learning models using supervised learning methods: Random Forests (RF), Nearest Shrunken Centroid (NSC), Gaussian Naive Bayes (G-NB), Logistic Regression (LogReg), and Stochastic Gradient Descent (SGD). A randomised search was used to perform five-fold cross-validation over a hyperparameter grid for each algorithm. Following this, the test set was used to evaluate each model with each feature set. This was done for each subtype and mutation considered in this work. Model performance was measured using metrics including Accuracy, Area Under the Receiver Operating Characteristics (AUROC), F1-Score, Matthews Correlation Coefficient (MCC), Precision, Recall, and Specificity.

Owing to the significant class imbalance in our dataset, we used MCC to determine the model for tasks requiring binary classification, such as predicting mutation status, and accuracy was selected as the metric for ranking models in other situations. Our public GitHub repository (*https://github.com/SP9144/GliomaMLClassifier*) contains all implementation code, including trained models and their weights.

D. Post Analysis

We performed gene set enrichment analysis (GSEA) and pathway analysis on the feature set corresponding to the best model for each case using Enrichr [16]. This enabled us to find novel or unique pathways and gene ontology terms that may be related to gliomas, as well as to confirm the biological relevance of the shortlisted genes against prior literature. To further investigate the association between each subtype or mutation and the prognostic group of list genes, we also carried out functional annotation clustering using Database for Annotation, Visualisation and Integrated Discovery (DAVID) [17]. This approach had a two-fold advantage. First, it verified that the genes we selected matched those previously discovered by other researchers and also assisted us in identifying previously unknown gene interactions.

IV. Discussion

The currency study presents a machine learning pipeline that applies a two-step feature extraction process and thorough training of the six supervised ML models using fivefold cross-validation over a hyperparameter grid. The goal is to classify a sample into subtypes of glioma and predict the mutation status of key genes of the assigned glioma subtype. Based on several performance metrics like MCC and accuracy, the optimal model is selected. We used MCC to handle a class imbalance in binary tasks as it considers all confusion matrix elements, offering a balanced metric for imbalanced data [18]. Additionally, Additionally, before determining significant mutations within each glioma subtype, a thorough analysis was conducted to obtain the important mutations present within glioblastoma subtypes. It included exploring each glioma's mutational landscape and performing survival analyses across the subtypes and selected genes. The feature sets corresponding to each best model were then used in enrichment analysis to verify biological relevance with previous findings.

Among the most altered genes in DG are IDH, TP53 and ATRX, with significant co-occurrence present among them ($p < 0.05$). The role of IDH as a key player in gliomas is confirmed by survival analysis, where the presence of a mutation in this gene led to prolonged survival. This also held for cases with co-occurrence of mutations of IDH, with TP53 and ATRX. We observed a survival advantage when the IDH mutation is present over the IDH wild-type irrespective of treatment received [19]. Patients with co-occurring mutations were observed to have higher survival rates, as per our findings. In particular, patients who had IDH mutations along with either TP53 or ATRX had much higher survival chances ($p < 0.0001$) than those who did not, implying that mutations that occur in conjunction demonstrate the potential to cooperate and improve a patient's survival outcomes. In the case of LGG, IDH mutation in the presence of 1p/19q codeletion has significantly longer survival than the LGG samples without these alterations. Most frequently found in oligodendrogliomas, the IDH mutations and codeletion are known to improve responsiveness to radiochemotherapy

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

[20]. The absence of samples with IDH wildtype and 1p/19q codeletion is explained by Labussi`ere et al. [21]. Survival analysis in EGFR and PTEN does not provide conclusive results. This is because GBM is a grade IV glioma, and the likelihood of survival is known to be exceedingly low, with less than 5% of patients surviving five years following diagnosis [22]. In addition to identifying key prognostic genes within each subtype, important transition and transversion trends, T>C changes and C>T transitions, which accounted for approximately 60% of the DG mutations. These mutational signatures confirm the validity of our classification method and add to the characteristics of the molecular features of gliomas.

Our models could accurately classify most samples using the ML models trained for DG, predicting outcomes with high values for all evaluation metrics. While one model found IDH mutations 98% of the time, the Random Forest model accurately classified samples into LGG and GBM groups. This demonstrates the accuracy of our approach in identifying significant genomic alterations and various glioma subtypes and also validates the strength of our feature sets. Despite the high performance in DG classification and mutation prediction, we observed an average performance (Accuracy: 65%) in the LGG subtype classification. This can be attributed to the historical challenges in the histopathological classification of these tumours. Additionally, since the classifier performed poorly in identifying samples as OA, accuracy in classifying the LGG subtype lowered. OA is also known as mixed glioma and develops from OD or A. This can be the underlying cause behind incorrect labelling of OA samples as A or OD. Adding to the challenges among subtypes is the poor prognosis of GBM [20]. Further, there was also a class imbalance caused by the lack of mutated samples (ranging from 27% to 30%) for EGFR and PTEN mutations. These challenges could be a possible explanation for the moderate performance of GBM when compared to DG.

Our findings from the enrichment and pathway analysis are consistent with those of earlier studies, indicating that the feature sets corresponding to the best-performing models had both significant biological relevance and predictive value. IDH, which is associated with oxidoreductase and upon mutation, isocitrate is first converted to α-KG, which is subsequently transformed to d-2-hydroxy-glutarate (d-2HG), or oxidoreductase [23]. Another prognostic gene is TP53, known as the guardian of the genome and a widely known tumour suppressor. Its loss can cause evasion of the cell’s immune response [24] and prevent apoptosis [25]. ATRX was recently reported to be present in corticotroph macroadenomas and carcinomas [26], and corticotroph tumours cause Cushing’s disease. Given that related terms recur, there may be an overlap in the prognostic feature sets of the co-occurring mutations and the individual mutations. Mutations in TERT, one of the top prognostic genes in DG, are common in gliomas and are exclusive to ATRX [27]. IDH, ATRX, and TERT promoter mutations in gliomas are correlated, according to a study by Ohba et al . [28], and there is a significant association between the lipid metabolism gene set and clinical features such as IDH mutation and 1p/19q codeletion [29]. A potential association between EGFR and cell adhesion has been identified [30], while PTEN controls glucose uptake and glucose transporter one expression in thyroid cancer cell s [31]. However, it is unclear if this is also true for gliomas. The enrichment of the HIF-1 signalling pathway is explained by the facilitation of HIF-1-mediated gene expression that results from PTEN loss [32].

Although the models presented in this study can make predictions with reasonable accuracy, there is always a concern associated with gene expression-based subtyping of heterogeneity over time and space. For instance, recent research has found that several GBM tumour subtypes can have distinct transcriptional subtypes [33]. The links between LGGs and GBMs that share genetic features like IDH mutation, or TERT promoter mutation status are still unclear based on current investigations. We can move toward an objective genome-based clinical classification, given that we have further explored these unknown links. By acquiring a thorough grasp of the subtype and mutations present in a patient using only the expression data of a select few genes, the current pipeline can also be used as a medical tool for prognosis. A patient’s targeted therapy or personalised medicine can subsequently be designed using the information from our pipeline. It can be improved by incorporating additional data modalities, such as imaging data, clinical data, etc., for prediction and can also be extended to treat other diseases.

V. CONCLUSION

The study presents an end-to-end pipeline for a detailed understanding of gliomas, including their subtypes and key mutations. Using both mutational patterns and patient clinical data, our models can identify the existence of key prognostic mutations and classify samples into their corresponding subtypes. Our pipeline, which combines mutation analysis, survival studies, and machine learning, provides a reliable tool for glioma classification. While performing well for most predictions, especially in distinguishing between LGG and GBM, the models show some limitations in classifying mixed gliomas. Adding more data modalities could improve the model’s performance and deepen our understanding of glioma biology. This work offers a practical and scalable tool for pathologists using precision medicine to treat glioma patients, with potential applications for other types of diseases.

Supplementary Information. The following supplementary files are provided with this study:

[1] **Supplementary File 1:** Complete mutational landscape analysis for each glioma subtype, including detailed co-occurrence patterns and Ti-Tv profiles.

[2] **Supplementary File 2:** Comprehensive survival analysis results of each glioma subtype and key mutation.

[3] **Supplementary File 3:** GO terms, KEGG pathways, and DAVID functional annotations enriched in prognostic genes across key glioma mutations.

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

REFERENCES

[1] Ostrom, Q.T., et al.: Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2013–2017. Neuro-Oncology 22(Supplement 1), 1–96 (2020)

[2] Louis, D.N., et al.: The 2021 who classification of tumors of the central nervous system: a summary. Neuro-Oncology 23(8), 1231–1251 (2021)

[3] Wen, P.Y., et al.: Glioblastoma in adults: a society for neuro-oncology (sno) and european society of neuro-oncology (eano) consensus review on current management and future directions. Neuro-Oncology 22(8), 1073–1113 (2020)

[4] Tan, A.C., et al.: Management of glioblastoma: State of the art and future directions. CA: A Cancer Journal for Clinicians 70(4), 299–312 (2020)

[5] Bent, M.J.: Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. Acta Neuropathologica 120(3), 297–304 (2010)

[6] Yan, H., et al.: Idh1 and idh2 mutations in gliomas. New England Journal of Medicine 360(8), 765–773 (2009)

[7] Chen, X., Zhang, T., Su, W., Dou, Z., Zhao, D., Jin, X., Lei, H., Wang, J., Xie, X., Cheng, B., Li, Q., Zhang, H., Di, C.: Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. Cell Death amp; Disease 13(11) (2022) https://doi.org/10.1038/s41419-022-05408-1

[8] Haase, S., Garcia-Fabiani, M.B., Carney, S., Altshuler, D., Nu´n˜ez, F.J., M´endez, F.M., Nu´n˜ez, F., Lowenstein, P.R., Castro, M.G.: Mutant atrx: uncovering a new therapeutic target for glioma. Expert Opinion on Therapeutic Targets 22(7), 599–613 (2018)

[9] Ceccarelli, M., et al.: Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 164(3), 550–563 (2016)

[10] Brat, D.J., et al.: Comprehensive, integrative genomic analysis of diffuse lowergrade gliomas. New England Journal of Medicine 372(26), 2481–2498 (2015)

[11] Verhaak, R.G., et al.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. Cancer Cell 17(1), 98–110 (2010)

[12] Brennan, C.W., et al.: The somatic genomic landscape of glioblastoma. Cell 155(2), 462–477 (2013)

[13] Aldape, K., et al.: Challenges to curing primary brain tumours. Nature Reviews

[14] Clinical Oncology 16(8), 509–520 (2019)

[15] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics 45(10), 1113–1120 (2013)

[16] Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., et al.: Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic acids research 44(8), 71–71 (2016)

[17] Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al.: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research 44(W1), 90–97 (2016)

[18] Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature protocols 4(1), 44–57 (2009)

[19] Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics 21(1), 1–13 (2020)

[20] Beiko, J., Suki, D., Hess, K.R., Fox, B.D., Cheung, V., Cabral, M., Shonka, N., Gilbert, M.R., Sawaya, R., Prabhu, S.S., Weinberg, J., Lang, F.F., Aldape, K.D., Sulman, E.P., Rao, G., McCutcheon, I.E., Cahill, D.P.: Idh1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection. Neuro-Oncology 16(1), 81–91 (2013)

[21] Cairncross, G., Wang, M., Shaw, E., Jenkins, R., Brachman, D., Buckner, J., Fink, K., Souhami, L., Laperriere, N., Curran, W., et al.: Phase iii trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of rtog 9402. Journal of Clinical Oncology 31(3), 337–343 (2013)

[22] Labussi`ere, M., Sanson, M., Idbaih, A., Delattre, J.-Y.: Combined analysis of tert, egfr, and idh status defines distinct prognostic glioblastoma classes. Neurology 74(17), 1408–1414 (2010)

[23] Ostrom, Q.T., Gittleman, H., Farah, P., Ondracek, A., Chen, Y., Wolinsky, Y., Stroup, N.E., Kruchko, C., Barnholtz-Sloan, J.S.: Cbtrus statistical report: Primary brain and central nervous system tumors diagnosed in the united states in 2006-2010. Neuro-oncology 15(suppl 2), 1–56 (2013)

[24] Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., et al.: Cancer-associated idh1 mutations produce 2-hydroxyglutarate. Nature 462(7274), 739–744 (2009)

[25] Aloni-Grinstein, R., Charni-Natan, M., Solomon, H., Rotter, V.: p53 and the viral connection: Back into the future ‡. Cancers 10(6), 178 (2018)

[26] Vaseva, A.V., Marchenko, N.D., Ji, K., Tsirka, S.E., Holzmann, S., Moll, U.M.: p53 opens the mitochondrial permeability transition pore to trigger necrosis. Cell 149(7), 1536–1548 (2012)

[27] Ebrahimi, A., Skardelly, M., Bonzheim, I., Ott, I., Mu¨hleisen, H., Eckert, F., Tabatabai, G., Schittenhelm, J.: Atrx immunostaining predicts idh and h3f3a status in gliomas. Acta Neuropathologica Communications 4(1) (2016)

[28] Heaphy, C.M., Wilde, R.F., Jiao, Y., Klein, A.P., Edil, B.H., Shi, C., Bettegowda, C., Rodriguez, F.J., Eberhart, C.G., Hebbar, S., et al.: Altered telomeres in tumors with atrx and daxx mutations. Science 333(6041), 425 (2011)

[29] Ohba, S., Kuwahara, K., Yamada, S., Abe, M., Hirose, Y.: Correlation between idh, atrx, and tert promoter mutations in glioma. Brain Tumor Pathology 37(2), 33–40 (2020) https://doi.org/10.1007/s10014-020-00360-4

[30] Pope, W.B., Prins, R.M., Albert Thomas, M., Nagarajan, R., Yen, K.E., Bittinger, M.A., Salamon, N., Chou, A.P., Yong, W.H., Soto, H., Wilson, N., Driggers, E., Jang, H.G., Su, S.M., Schenkein, D.P., Lai, A., Cloughesy, T.F., Kornblum, H.I., Wu, H., Fantin, V.R., Liau, L.M.: Non-invasive detection of 2hydroxyglutarate and other metabolites in idh1 mutant glioma

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)

patients using magnetic resonance spectroscopy. Journal of Neuro-Oncology 107(1), 197–205 (2011)

[31] Larsen, A.B., Pedersen, M.W., Stockhausen, M.-T., Grandal, M.V., Van Deurs, B., Poulsen, H.S.: Epidermal growth factor induces selective phosphorylation of the epha2 receptor. Journal of Biological Chemistry 282(19), 14122–14128 (2007)

[32] Morani, F., Phadngam, S., Follo, C., Titone, R., Aimaretti, G., Galetto, A., Isidoro, C.: Pten regulates plasma membrane expression of glucose transporter 1 and glucose uptake in thyroid cancer cells. Journal of Molecular Endocrinology 53(2), 247–258 (2014)

[33] Zundel, W., Schindler, C., Haas-Kogan, D., Koong, A., Kaper, F., Chen, E., Gottschalk, A.R., Ryan, H.E., Johnson, R.S., Jefferson, A.B., et al.: Loss of pten facilitates hif-1-mediated gene expression. Genes & Development 14(4), 391–396 (2000)

[34] Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al.: Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344(6190), 1396–1401 (2014)

(This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License citing the original author and source)