Classifying Gliomas as LGG vs GBM using Machine Learning Models

Lavanya Pushparaj

***Abstract*—Gliomas, classified into Low-Grade Gliomas (LGGs) and Glioblastoma Multiforme (GBMs), are brain tumors characterized by the overgrowth of glial cells. This study aimed to distinguish between LGGs and GBMs using genetic profiles, focusing on mutations such as IDH1, IDH2, TP53, and others. Three machine learning models—Support Vector Machine (SVM), Random Forest, and XGBoost—were utilized to analyze 839 instances with 23 features from the ”Glioma Grading Clinical and Mutation Features” dataset. The SVM model showed superior performance, achieving an accuracy of 0.86 with high precision and recall, closely followed by XGBoost. The findings suggest that specific gene mutations, particularly IDH1 and IDH2, are critical in differentiating between LGG and GBM, which could enhance diagnostic accuracy and potentially guide treatment strategies. Further research is needed to explore the underlying mechanisms of these mutations and develop targeted interventions for glioma management.**

I. Introduction

Gliomas are a form of brain tumor that results from an excess accumulation of glial cells in the central nervous system[1]. Glial cells are essential in supporting, connecting, and protecting neurons[2]. While neurons are commonly known for nervous system functions, glial cells are equally important, if not more. They mediate and support the neurons, allowing them to function by mediating neuronal function and signaling. Gliomas start with the overgrowth of brain or spinal cord glial cells[3]. While these glial cells may look and act like normal cells, the fact that they are in abundance is the main issue. Gliomas are often categorized into Low-Grade Gliomas (LGG) and Glioblastoma Multiforme (GBM). LGGs are benign neuroepithelial tumors that are noncancerous, grow slower, and do not spread into other body parts. LGGs encompass both grade 1 and grade 2 tumors and may be caused by genetic or environmental factors. Mutations in tumor suppressor protein 53 (p53), phosphatase and tensin homolog (PTEN), and epidermal growth factor receptor (EGFR) are often seen in the pathophysiology of patients with LGG.[4] Over time, LGGs can transform into GBMs. GBMs are malignant tumors that include astrocytic tumors (astrocytoma, anaplastic astrocytoma, and glioblastoma), oligodendrogliomas, ependymomas, and mixed gliomas. GBMs include grades 3 and 4, which are significantly more aggressive than the lower grades. GBMs are more common than LGGs and are more challenging to cure due to the lack of knowledge on the etiology of this condition. Mutations of IDH1/2, TP53, and ATRX have been commonly associated with the presence of GBMs [5]. My research aims to identify a way to predict the presence of LGG vs. GBM in an individual by studying the influence of a combination of specific mutations. Three machine learning models were built, and the most effective one was identified.

II. Background

A. Genes

The gene mutations under investigation in this study include IDH1, IDH2, TP53, FAT4, ATRX, PTEN, EGFR, CIC, MUC16, PIK3CA, NF1, PIK3R1, PDGFRA, FUBP1, RB1, NOTCH1, BCOR, CSMD3, SMARCA4, and GRIN2A. Mutations in IDH1 and IDH2 alter enzyme function, leading to the production of 2-hydroxyglutarate and increased methylation in gliomas[6,7,8]. TP53 encodes the tumor suppressor p53, while PTEN and RB1 also produce proteins that suppress tumors[9]. EGFR, the epidermal growth factor receptor, regulates cell division and survival. FAT4 is involved in tissue protein formation, and ATRX maintains genomic stability[10]. Further, CIC acts as a transcriptional repressor, MUC16 serves as a cancer antigen, and PIK3CA regulates cell functions [11]. NF1, encoding neurofibromin, mediates neuronal signaling. PIK3R1 and PDGFRA are essential in cell growth and signaling pathways. FUBP1 functions as a regulatory molecule, NOTCH1 promotes cell growth, and BCOR, involved in immune system regulation, produces BCL6[12]. CSMD3's role is less defined but may involve macrophage promotion. SMARCA4 is crucial for DNA repair and regulation, and GRIN2A affects NMDA receptors and synaptic functions[13]. Understanding the specific patterns of these mutations can help differentiate between Low-Grade Gliomas (LGGs) and Glioblastoma Multiforme (GBMs), aiming to enhance prediction and early diagnosis for patients exhibiting these genetic profiles.

B. Models

Three classification models of importance for this research include Support Vector Machine (SVM), Random Forest, and XGBoost. SVMs utilize a decision boundary to separate the labels while maximizing the margin. Random Forest is a group of multiple decision trees combined into a single model to formulate a prediction between the different classes of the label. XGBoost is an ensemble learning method and iteratively uses decision trees to make its final prediction. All three models have been previously used to represent and make predictions for cancer diagnoses in the past. [14, 15, 16]

III. Related Work

Based on prior research, mutations in IDH1 and IDH2 are most common in gliomas. It is commonly seen in patients with grade II or grade III tumors, with a prevalence of greater than 80%. Grades II and III are categorized as LGGs. This mutation is not as common in primary GBMs and can be seen with the significantly lower prevalence of 3.7% diagnoses.[6,7,8] Additionally, machine learning methods have been utilized to classify gliomas using MR spectroscopy. In this case, the gliomas were categorized as benign or malignant. The LGGs were categorized as benign, and the GBMs were categorized as GBMs. Their MR spectroscopy images were used to run the models and create the classifications.[17] Additionally, more research was done using biological markers to predict the glioma grade. This research indicated co-enrichment between glioma grade-related annotations, indicating connections between the pathways and genes that result in LGGs vs GBMs. Research into gliomas is relatively small and still has much room for growth and further inquiry.[18]

IV. Methodology

A. The Data

The dataset from the UC Irvine Machine Learning Repository was called ”Glioma Grading Clinical and Mutation Features.” This dataset consists of 839 instances and 23 features and includes a target label. Since a target label, ’Grade’ (the two potential labels were LGG and GBM), was included, a classification task was performed. The features include: Project, Case_ID, Gender, Age_at_diagnosis, Primary_Diagnosis, Race, IDH1, TP53, FAT4, ATRX, PTEN, EGFR, CIC, MUC16, PIK3CA, NF1, PIK3R1, PDGFRA, FUBP1, RB1, NOTCH1, BCOR, CSMD3, SMARCA4, GRIN2A, and IDH2.

B. Preprocessing

Each feature was determined to be either categorical or numerical. The categorical columns underwent label encoding. For example, the values for all the genes include: ”NOT_MUTATED” or ”MUTATED.” The ’Age_at_diagnosis’ had values such as ”59 years 32 days,” and the only value retrieved and stored was the value for years. Any rows with missing values were dropped. The gender and race class were analyzed for any values of ’–,” which indicates that the individual did not prefer to specify. Bar plots were then utilized to visualize the data and to see the spread of the data.

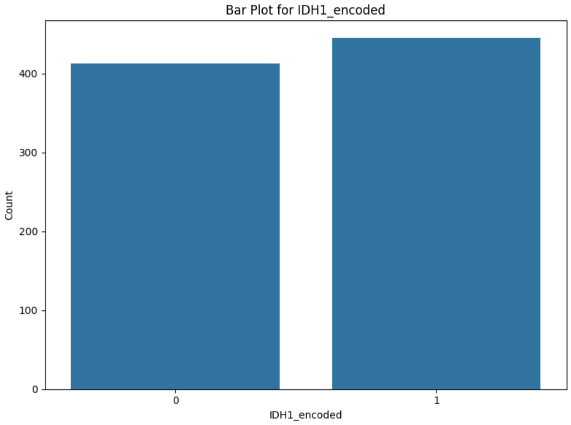

Fig. 1. Bar Plot of IDH1 of mutated vs not mutated

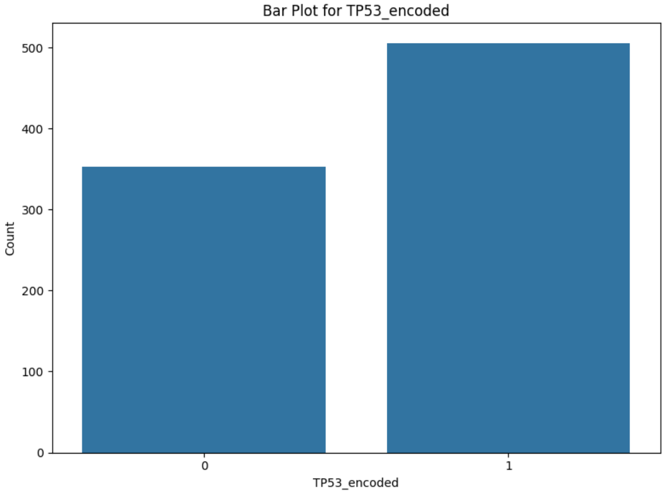

Fig. 2. Bar plot of TP53 of mutated vs not mutated

The features and the target were specified, and then the data was split into a train-test with a 70-30 split. The features were also normalized and were finally ready for model training. It was only after the first implementation that it was discovered that the feature ’Primary_Diagnosis_encoded’ also needed to be dropped. This was because the model was overfitting when this feature was left in the dataset. After further investigation, this overfitting was because the diagnosis indicated what type of glioma it was, which would directly correlate to whether it was an LGG or a GBM. A correlation heatmap was also constructed to visualize the relationship between each variable. Two variables had to be dropped, ’Case_ID_encoded’ and ’Project_encoded,’ due to a correlation value above 0.8 (0.8407524435225795).

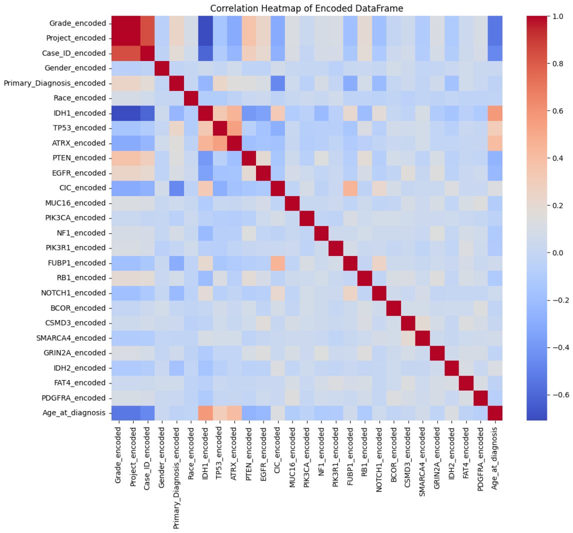

Fig. 3. Correlation Heat Map

C. Models

Then, the individual SVM, random forest, and XGBoost models were built upon the preprocessed data. The accuracy and the classification report were printed to check how well each model performed. For each of the models, hyperparameters were specified. For the SVM model, the kernel used was linear to allow for linear separation. For the random forest model, N_estimators were set to 50, indicating the number of decision trees utilized, and the random_state was set to 42 for reproducibility. XGBoost included multiple hyperparameters that needed to be tuned for a better model and data representation. The use_label_encoder was set to False because label encoding was already utilized during preprocessing. The eval_metric was set to 'logloss,' and the max_depth was set to 6, indicating the max depth of a tree. The colsample_bytree was set to 0.8, indicating that only 80% of the feature columns were used for each decision tree. Finally, the gamma was set to 0.1 to indicate minimum loss reduction.

V. Result

A. SVM

The SVM resulted in an accuracy of 0.86. For class 0 (LGG), the precision was 0.78, which means that 78% of the instances predicted as class 0 are class 0; the recall was 0.94, which means that the model identifies 94% of all actual class 0 instances; the F1-Score was 0.85. For class 1 (GBM), the precision was 0.95, which means that 95% of the instances predicted as class 1 are class 1; the recall is 0.81, and this means that the model Identifies 81% of all actual class 1 instances; and the F1-Score was 0.87.

B. Random Forest

The random forest resulted in an accuracy of 0.83. For class 0 (LGG), the precision was 0.79, which means that 79% of the instances predicted as class 0 are class 0; the recall was 0.81, which means that the model identifies 81% of all actual class 0 instances; the F1-Score was 0.80. For class 1 (GBM), the precision was 0.86, which means that 86% of the instances predicted as class 1 are class 1; the recall is 0.85, and this means that the model Identifies 85% of all actual class 1 instances; and the F1-Score was 0.85.

C. XGBoost

The XGBoost resulted in an accuracy of 0.86. For class 0 (LGG), the precision was 0.81, which means that 81% of the instances predicted as class 0 are class 0; the recall was 0.86, which means that the model identifies 86% of all actual class 0 instances; the F1-Score was 0.84. For class 1 (GBM), the precision was 0.89, which means that 89% of the instances predicted as class 1 are class 1; the recall is 0.85, and this means that the model Identifies 85% of all actual class 1 instances; and the F1-Score was 0.87.

D. Comparison

In comparing the performance of Support Vector Machine (SVM), Random Forest, and XGBoost classification models, both SVM and XGBoost demonstrate the highest overall accuracy at 0.86. SVM has a high precision of 0.95 for class 1 (GBM) and the highest recall of 0.94 for class 0 (LGG), indicating its ability to classify both tumor types accurately. XGBoost also has high metrics, with a high precision of 0.81 for class 0 and high F1 scores, particularly for class 1. Random Forest has lower values for all the metrics. Overall, SVM is the better model because of its balanced performance, high precision, and recall. However, XGBoost could also be a reliable model. A graph was created to see the importance of the features of this model. As can be seen, IDH1 had the highest importance, followed by IDH2 and FUBP1. This matches with what previous findings have indicated, that IDH1 and IDH2 heavily influence gliomas, specifically in categorizing the two types of gliomas.

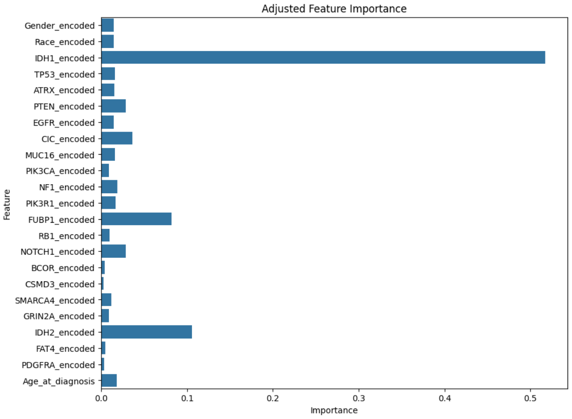

Fig. 4. Feature Importance

VI. Conclusion

Further research must be done to understand these mutations' accurate functions and implications and how they can be reversed or managed to help formulate a cure. GBMs require more understanding; for now, little more information has yet to be ascertained on how mutations and other environmental factors lead to the aggressive form of tumor spotted in many patients. This type of predictive modeling can also help provide patients with a diagnosis[19]. Sometimes, doctors may miss a mutation while manually reviewing a patient's charts. However, machine learning models can track multiple features simultaneously to provide a more unbiased diagnosis[20]. Diagnosing an LGG correctly can prevent its morphing and escalating to a GBM. Additionally, patients with GBM can be correctly identified and provided with the necessary care and treatments.

VII. References

1) Ohgaki, H., Kleihues, P. Epidemiology and etiology of gliomas. Acta Neuropathol 109, 93–108 (2005). https://doi.org/10.1007/s00401-005-0991-y
2) Purves D, Augustine GJ, Fitzpatrick D, et al., editors. Neuroscience. 2nd edition. Sunderland (MA): Sinauer Associates; 2001. Neuroglial Cells. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10869/
3) Mesfin FB, Al-Dhahir MA. Gliomas. [Updated 2023 May 20]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK441874/

4) Aiman W, Gasalberti DP, Rayi A. Low-Grade Gliomas. [Updated 2023 May 6]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560668/
5) Hanif, F., Muzaffar, K., Perveen, K., Malhi, S. M., & Simjee, S.hU. (2017). Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. Asian Pacific journal of cancer prevention : APJCP, 18(1), 3–9. https://doi.org/10.22034/APJCP.2017.18.1.3
6) Cohen, A. L., Holmen, S. L., & Colman, H. (2013). IDH1 and IDH2 mutations in gliomas. Current neurology and neuroscience reports, 13(5), 345. https://doi.org/10.1007/s11910-013-0345-4
7) Han, S., Liu, Y., Cai, S.J. et al. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. Br J Cancer 122, 1580–1589 (2020). https://doi.org/10.1038/s41416-020-0814-x
8) Yan H, Parsons DW, Jin G, et al.: IDH1 and IDH2 mutations in gliomas. N Engl J Med. 2009, 360:765-73. 10.1056/NEJMoa0808710
9) Noor, H., Briggs, N. E., McDonald, K. L., Holst, J., & Vittorio, O. (2021). TP53 Mutation Is a Prognostic Factor in Lower Grade Glioma and May Influence Chemotherapy Efficacy. Cancers, 13(21), 5362. https://doi.org/10.3390/cancers13215362
10) Nandakumar, P., Mansouri, A., & Das, S. (2017). The Role of ATRX in Glioma Biology. Frontiers in oncology, 7, 236. https://doi.org/10.3389/fonc.2017.00236
11) Brito, C., Tomás, A., Azevedo, A., Esteves, S., Mafra, M., Roque, L., & Pojo, M. (2022). PIK3CA Mutations in Diffuse Gliomas: An Update on Molecular Stratification, Prognosis, Recurrence, and Aggressiveness. Clinical Medicine Insights. Oncology, 16, 11795549211068804. https://doi.org/10.1177/11795549211068804
12) Jiao, Y., Killela, P. J., Reitman, Z. J., Rasheed, A. B., Heaphy, C. M., de Wilde, R. F., Rodriguez, F. J., Rosemberg, S., Oba-Shinjo, S. M., Nagahashi Marie, S. K., Bettegowda, C., Agrawal, N., Lipp, E., Pirozzi, C., Lopez, G., He, Y., Friedman, H., Friedman, A. H., Riggins, G. J., Holdhoff, M., … Yan, H. (2012). Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas. Oncotarget, 3(7), 709–722. https://doi.org/10.18632/oncotarget.588
13) Chen, C. C. L., Andrade, A. F., & Jabado, N. (2023). SMARCA4 vulnerability in H3K27M midline glioma: A silver bullet for a lethal disease. Molecular cell, 83(2), 163–164. https://doi.org/10.1016/j.molcel.2022.12.028
14) Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer genomics & proteomics, 15(1), 41-51.
15) Ram, M., Najafi, A., & Shakeri, M. T. (2017). Classification and biomarker genes selection for cancer gene expression data using random forest. Iranian journal of pathology, 12(4), 339.
16) Liew, X. Y., Hameed, N., & Clos, J. (2021). An investigation of XGBoost-based algorithm for breast cancer classification. Machine Learning with Applications, 6, 100154.
17) Ranjith, G., Parvathy, R., Vikas, V., Chandrasekharan, K., & Nair, S. (2015). Machine learning methods for the classification of gliomas: Initial results using features extracted from MR spectroscopy. The neuroradiology journal, 28(2), 106–111. https://doi.org/10.1177/1971400915576637
18) Garbulowski, M., Smolinska, K., Çabuk, U., Yones, S. A., Celli, L., Yaz, E. N., Barrenäs, F., Diamanti, K., Wadelius, C., & Komorowski, J. (2022). Machine Learning-Based Analysis of Glioma Grades Reveals Co-Enrichment. Cancers, 14(4), 1014. https://doi.org/10.3390/cancers14041014
19) Yang C. C. (2022). Explainable Artificial Intelligence for Predictive Modeling in Healthcare. Journal of healthcare informatics research, 6(2), 228–239. https://doi.org/10.1007/s41666-022-00114-1
20) Vogenberg F. R. (2009). Predictive and prognostic models: implications for healthcare decision-making in a modern recession. American health & drug benefits, 2(6), 218–222.