Exploring Clinical and Molecular Features for Glioma Grading Using TCGA-LGG and TCGA-GBM Datasets

Piyush Dubey
Student B.Tech CSE, LPU, Phagwara, Punjab, India.

Dr. Dhanpratap
Assistant Professor , LPU, Phagwara, Punjab, India.

Abstract : The grading of gliomas based on clinical and mutational characteristics is covered in this article. Glioma is the easily found brain tumour, which can be classified as GBM or LGG depending on histology and imaging standards. However, the grading method also takes into account clinical and molecular/mutational factors. The most common clinical traits of gliomas, including headaches, seizures, cognitive decline, motor deficits, language difficulties, and personality abnormalities, are highlighted in the article. IDH mutations, TP53 mutations, EGFR mutations, BRAF mutations, NF1 mutations, chromosome 10 loss, 7 gain and 1p/19q loss are the only a few examples of the genetic and chromosomal abnormalities that are mentioned. The 20 most frequently altered genes and three clinical traits from the TCGA-Low Grade Glioma and TCGA- Glioblastoma Multiforme are discussed in my paper.

1. Introduction:

The Glioma is the easily identified and occurring brain tumour. It can be categorised as GBM (Glioblastoma Multiforme) or LGG (Lower-Grade Glioma) based on the histology and images . Clinical mutation and molecular criteria are also very important for the grading of the tumour. It is costly to conduct molecular studies to precisely check glioma patients. I have used the dataset which contains twenty mutated and three clinical feature from Low

Grade Glioma and Glioblastoma Multiforme . A patient's LGG or GBM status is predicted based on a variety of clinical mutational features and molecular. The aim then requiring to select the suitable clinical characteristics and group of the mutant genes for the glioma tumour grading so that it may help in enhancing productivity and reducing the cost.

Background :

The Clinical characteristics of glioma refer to the signs, symptoms, and risk factors that may be present in persons with this type of brain tumour. The following are some of the most prevalent clinical traits of glioma:

Headaches: Patients with gliomas frequently have them, and their intensity and frequency can vary.

Seizures: Patients with high-grade tumours are more likely to experience seizures as a glioma symptom.

Cognitive impairment: Patients may notice changes in their memory, concentration, or other cognitive abilities depending on where the tumour is located.

Motor deficits: Patients with gliomas may develop coordination issues, limb numbness or weakness, or other motor deficiencies.

Gliomas that are close to the optic nerve or other visual pathways might cause patients to experience visual abnormalities.

Language problems: Gliomas in specific brain regions can impair a patient's capacity to communicate, leading to language problems.

Personality changes: Patients may experience changes in their personalities, behaviours, or moods, depending on the location and size of the tumour.

In addition to these symptoms, glioma has been linked to a number of risk factors, such as age, a family history of brain tumours, and radiation exposure.

It is important to emphasise that the patient's general health and other factors, the location, size, and grade of the tumour, as well as other factors, can all have an impact on these clinical characteristics. Imaging techniques like computed tomography (CT scan)or magnetic resonance imaging (MRI scans) are frequently used to diagnose and assess gliomas clinically.

Genetic and chromosomal abnormalities that have been found in various kinds of brain tumours are referred to as glioma mutation characteristics. For the purpose of glioma diagnosis and treatment, it is crucial to comprehend the genetic mutations and chromosomal abnormalities connected with this condition.

Glioma is always being confined to several mutation which includes:

IDH mutations: The most frequent genetic changes identified in gliomas are those in the IDH1 or IDH2 genes. Longer survival and a better prognosis are linked to certain mutations.

TP53 mutations: The TP53 gene is frequently mutated in gliomas and it is usually linked to a aggressive tumours

EGFR mutations: A fraction of gliomas have EGFR gene mutations, it is usually linked to a aggressive tumours .

BRAF mutations: BRAF gene mutations have been found in some low-grade gliomas and are linked to a better prognosis.

NF1 mutations: Gliomas that develop in people with neurofibromatosis type 1 are linked to mutations in the NF1 gene.

Chromosomal abnormalities have also been seen in gliomas in addition to genetic alterations. Gains or losses of entire chromosomes or specific chromosomal regions are examples of these disorders. The following are some of the most typical chromosomal abnormalities in glioma:

Chromosome 10 loss: Chromosome 10 loss is a frequent finding in gliomas and is linked to a more aggressive tumour phenotype.

Chromosome 7 gain: Chromosome 7 gain is another mutation that frequently occurs in gliomas and is linked to an aggressive tumour phenotype.

Chromosome 1p/19q loss: Chromosome 1p/19q loss is found in oligodendrogliomas, a subset of gliomas, and is related to a better prognosis.

The diagnosis and management of glioma can be aided by an understanding of its mutational characteristics. Patients who have IDH mutations, for instance, may benefit from tailored medicines that take advantage of this mutation. Additionally, chromosomal anomalies can be utilised to divide gliomas into various subtypes, which can aid in determining the best course of treatment.

1.1 About Data:

The Low grade glioma and Glioblastoma Multiforme brain glioma shows twenty most frequently mutated genes , three clinical characteristics are taken into consideration in this dataset.

The pre processed and arranged CSV dataset has 24 fields per record. Each field is separated by a comma, and each record is separated by a newline. Clinical considerations include gender, age upon diagnosis, and racial characteristics. These traits may be mutated or not based on the id of TGCA

See the attributes here:

1.Gender: 0 for males and 1 for females.

2. Age_at diagnosis: Age at diagnosis plus days estimated

3.Race : For Race

(a.0 = white,

b.1 = black or African American,

c.2 = Asian, and

d.3 = American Indian or Alaskan Native)

4.IDH1: isocitrate dehydrogenase (NADP(+))1 (0 = NOT_MUTATED; 1= MUTATED)

5.TP53: Tumour protein p53 (MUTATED = 1; NOT_MUTATED = 0).

6.ATRX: ATRX chromatin remodeler (MUTATED = 1; NOT_MUTATED = 0).

7.PTEN: homolog of the enzymes phosphatase and tensin (0 = NOT_MUTATED; 1 = MUTATED)

8.EGFR: the epidermal growth factor receptor (MUTATED = 1; NOT_MUTATED = 0).

9.CIC: Capicua transcriptional repressor (0: NOT_MUTATED; 1: MUTATED)

10.MUC16: mucin 16, related with the cell surface (0 = NOT_MUTATED; 1 = MUTATED).

11. Phosphatidylinositol-4,5-bisphosphate 3-kinase Catalytic Subunit Alpha (PIK3CA) (0 = NOT_MUTATED; 1 = MUTATED)

12.NF1: Neurofibromin 1 (MUTATED = 1; NOT_MUTATED = 0).

13. Phosphoinositide-3-kinase regulatory subunit 1 (PIK3R1) (0 = NOT_MUTATED; 1 = MUTATED).

14. Far upstream element binding protein 1 (FUBP1) : (0 = NOT_MUTATED, 1 = MUTATED).

15. RB1 is the RB transcriptional corepressor number one (0 = NOT_MUTATED; 1 = MUTATED).

16.NOTCH1: Narrowband Transmitter Receptor 1 (0 = NOT_MUTATED; 1 = MUTATED)

17. BCOR= BCL6 corepressor (0 = NOT_MUTATED; 1 = MUTATED).

18.CSMD3: Multiple domains for CUB and sushi 3 (NOT_MUTATED = 0; MUTATED = 1)

19.SMARCA4: member of the subfamily an of the SWI/SNF related, matrix associated, actin dependent regulator of chromatin genes. (1 = MUTATED; 0 = NOT_MUTATED)

20. NMDA type glutamate ionotropic receptor subunit 2A. (1 = MUTATED; 0 = NOT_MUTATED)

21. Isocitrate dehydrogenase (NADP(+)) is number 21. 2 (NOT_MUTATED = 0; MUTATED = 1)

22.FAT4 refers to the FAT atypical cadherin 4 (MUTATED = 1; NOT_MUTATED = 0).

23.Platelet-derived growth factor receptor alpha (PDGFRA) (0 = NOT_MUTATED; 1 = MUTATED).

Important:

The following information is provided for the class label:

glioma grade class (1 = GBM; 0 = LGG)

2. Data Pre-processing:

The pre-processed dataset contains 23 instances where the Age ,Gender, Race . Values are hyphen or 'not reported'. These examples, the Project, Case ID, and Primary Diagnosis columns from the original dataset file were taken out to create the pre-processed dataset file.

Age values were transformed from string to conti-nuous values during the pre processing stage by adding information to dataset.

	Grade	Project	Case_ID	Gender	Age_at_diagnosis	Primary_Diagnosis	Race	IDH1	TP53	ATRX	...	FUBP1	RB1	NOTCH1	
0	LGG	TCGA-LGG	TCGA-DU-8164	Male	51 years 108 days	Oligodendroglioma, NOS	white	MUTATED	NOT_MUTATED	NOT_MUTATED	...	MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
1	LGG	TCGA-LGG	TCGA-QH-A6CY	Male	38 years 261 days	Mixed glioma	white	MUTATED	NOT_MUTATED	NOT_MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
2	LGG	TCGA-LGG	TCGA-HW-A5KM	Male	35 years 62 days	Astrocytoma, NOS	white	MUTATED	MUTATED	MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
3	LGG	TCGA-LGG	TCGA-E1-A7YE	Female	32 years 283 days	Astrocytoma, anaplastic	white	MUTATED	MUTATED	MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
4	LGG	TCGA-LGG	TCGA-S9-A6WG	Male	31 years 187 days	Astrocytoma, anaplastic	white	MUTATED	MUTATED	MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
857	GBM	TCGA-GBM	TCGA-19-5959	Female	77 years 325 days	Glioblastoma	white	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
858	GBM	TCGA-GBM	TCGA-16-0846	Male	85 years 65 days	Glioblastoma	white	NOT_MUTATED	MUTATED	NOT_MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
859	GBM	TCGA-GBM	TCGA-28-1746	Female	77 years 178 days	Glioblastoma	white	NOT_MUTATED	MUTATED	NOT_MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M
860	GBM	TCGA-GBM	TCGA-32-2491	Male	63 years 121 days	Glioblastoma	white	NOT_MUTATED	MUTATED	NOT_MUTATED	...	NOT_MUTATED	MUTATED	NOT_MUTATED	NOT_M
861	GBM	TCGA-GBM	TCGA-06-2557	Male	76 years 221 days	Glioblastoma	black or african american	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	...	NOT_MUTATED	NOT_MUTATED	NOT_MUTATED	NOT_M

```
Grade                0
Project              0
Case_ID              0
Gender               0
Age_at_diagnosis     0
Primary_Diagnosis    0
Race                 0
IDH1                 0
TP53                 0
ATRX                 0
PTEN                 0
EGFR                 0
CIC                  0
MUC16                0
PIK3CA               0
NF1                  0
PIK3R1               0
FUBP1                0
RB1                  0
NOTCH1               0
BCOR                 0
CSMD3                0
SMARCA4              0
GRIN2A               0
IDH2                 0
FAT4                 0
PDGFRA               0
dtype: int64
```

2.1 Dataset Characteristics:

The data is Tabular in nature and of Multivariate form.

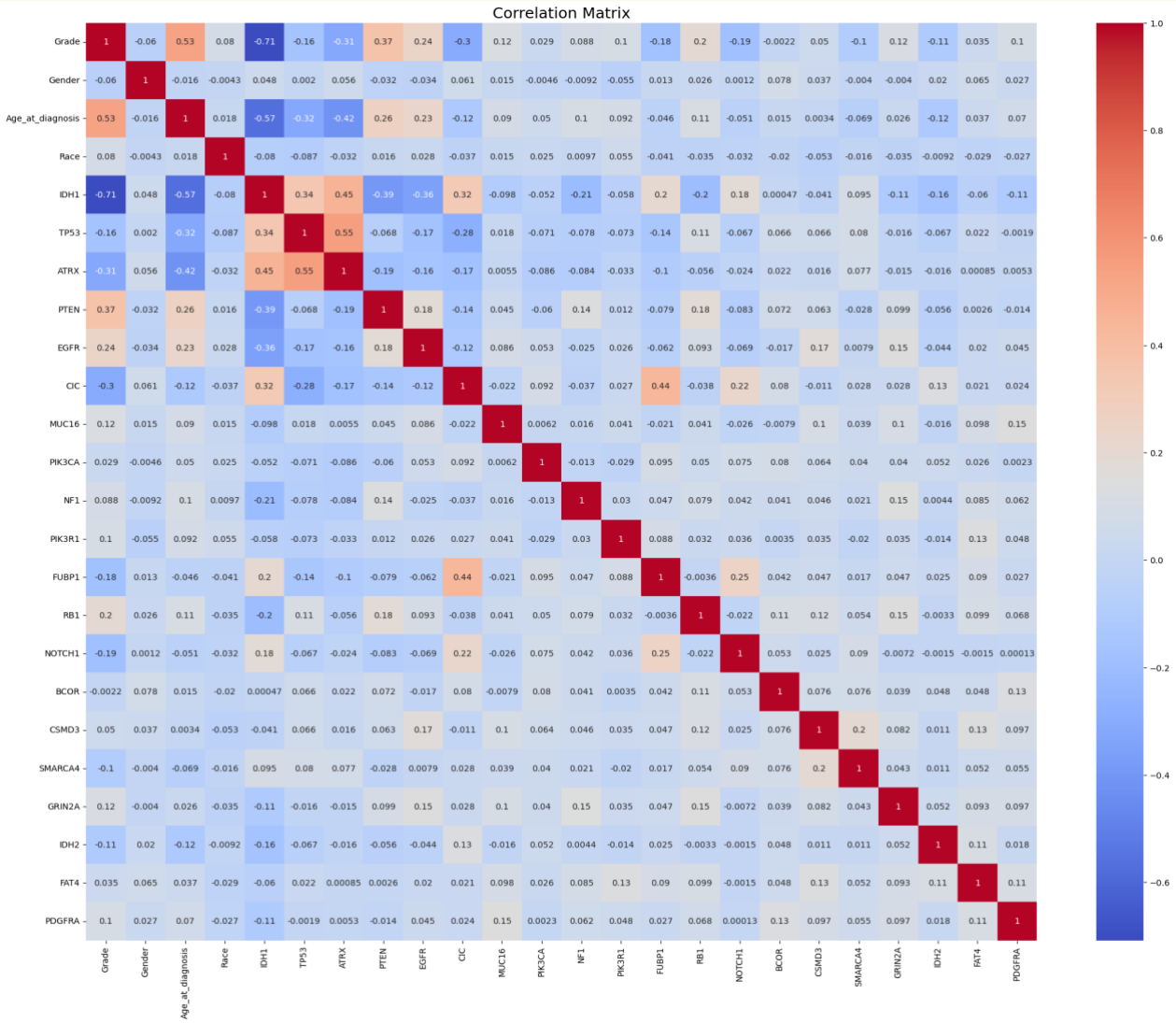
2.2Attribute Type:

Real, Categorical, Integer

3. Model Selection and Training:

Random Forest, Ensemble, K Neighbour Classifier, Decision Trees, and Support Vector Machine with various arguments were among the machine learning methods that were assessed. To improve each model's hyperparameters, grid search was performed. Metrics like precision, accuracy, F1 score , recall were used to check the models.

To scale features to a range between [-1, 1], machine learning uses a form of data normalisation algorithm called MaxAbsScaler. It operates by dividing each feature in the dataset by its highest absolute value. When the data contains outliers and we want to maintain the data's sparsity, this form of scaling is helpful. We aim to plot the heatmap of the corelation matrix after scaling.



Now , Using **Random Forest Classifier** , we trained our model and generated the output with the accuracy of 87%

```
Best Parameters:  {'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_
Best Score:  0.8722439519049688
Accuracy score:  0.8809523809523809
```

Using **Ensemble** , after training the data and finding the hyperparamaters best suited , we get an accuracy of 85.71%

```
Best hyperparameters:  {'learning_rate': 0.1, 'n_estimators': 50}
Accuracy score:  0.8571428571428571
```

Using **Decision Tree Classifier** , after training the data, defining the hyperparameter to tune , we got accuracy approx. 85.1%

```
Best hyperparameters: {'max_depth': 3, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 2}
Accuracy: 0.8511904761904762
```

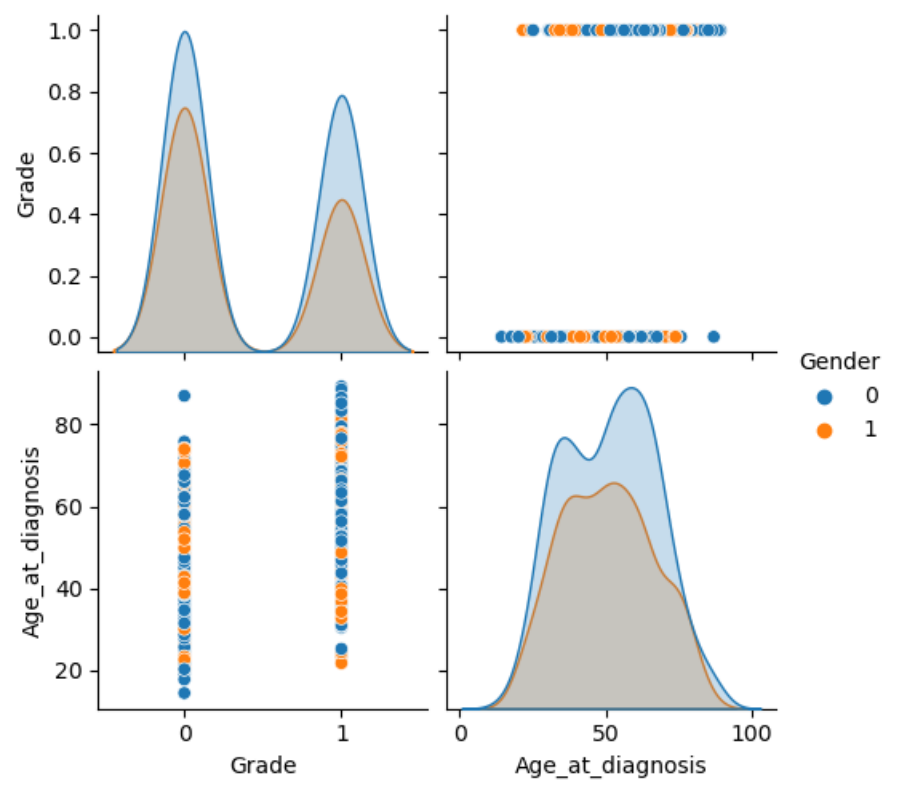
Using K Neighbour Classifier , after training the data and hyperparameter tuning , the model performed at 86.5%

```
Best hyperparameters:  {'n_neighbors': 9, 'p': 1, 'weights': 'uniform'}
Best score:  0.8658706467661691
[[74 15]
 [12 67]]
              precision    recall  f1-score   support

           0       0.86      0.83      0.85        89
           1       0.82      0.85      0.83        79

    accuracy                           0.84       168
   macro avg       0.84      0.84      0.84       168
weighted avg       0.84      0.84      0.84       168
```

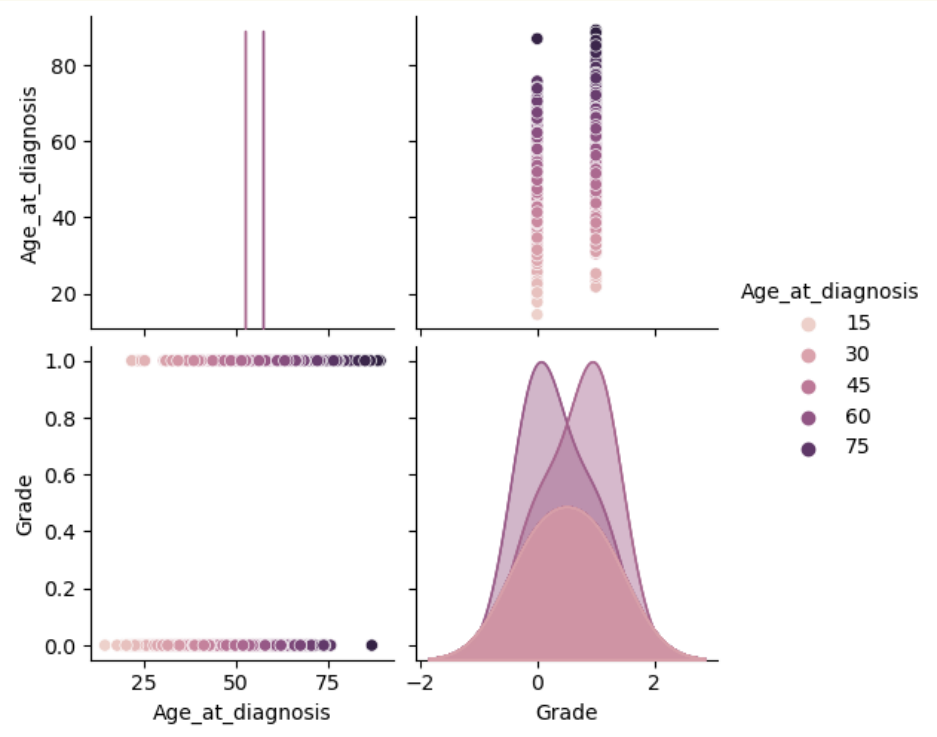
Using Gradient Boosting Classifier , after training the data and hyperparameter tuning , I received approx. 84.5% accuracy.

```
Accuracy: 0.8571428571428571
[[70 19]
 [ 5 74]]
              precision    recall  f1-score   support

           0       0.93      0.79      0.85        89
           1       0.80      0.94      0.86        79

    accuracy                           0.86       168
   macro avg       0.86      0.86      0.86       168
weighted avg       0.87      0.86      0.86       168
```

The Pair Plots generated include:



And.



4. Model Training:

The selected SVM, Logistic Regression, KNN, Decision Tree Classifier model with a linear kernel was trained on the training set with the optimal hyperparameters. The model was then used to make predictions on the testing set.

5. Model Improvement:

To improve the model's interpretability, LIME (Local Interpretable Model-agnostic Explanations) was used to explain the predictions. The top 5 features that contributed to the model's prediction were identified.

6. Conclusion:

In conclusion, a machine learning model was successfully developed to classify the patients with glioma into two categories based on clinical mutation & molecular features. The Random Forest model with a linear kernel achieved the best performance with an accuracy of 87%. The model's interpretability was improved by using LIME to explain the predictions. The top 5 features that contributed to the model's prediction were identified, providing valuable insights into the underlying biological mechanisms of gliomas.

The examination of both clinical and molecular/mutational criteria is necessary for the difficult process of glioma grading. Patients with gliomas experience symptoms that can be utilised for diagnosis and grading, such as headaches, seizures, cognitive decline, motor deficits, visual problems, language difficulty, and personality changes. On the other hand, chromosomal abnormalities such as loss of chromosome 10, gain of chromosome 7, loss of chromosome 1p/19q, amplification of the EGFR gene, and loss of the CDKN2A/B locus are also significant for glioma grading. These molecular characteristics include IDH mutations, TP53 mutations, EGFR mutations, BRAF mutations, and chromosomal abnormalities.

The accuracy of diagnosis, cost savings, and patient outcomes can all be enhanced by applying machine learning models to predict the grade of gliomas based on clinical and molecular/mutation data. The most frequently altered 20 genes and 3 clinical characteristics from the TCGA-LGG and TCGA-GBM brain glioma projects were included in the dataset examined in this investigation. This work offers important insights for enhancing the grading of gliomas by determining the ideal subset of mutation genes and clinical characteristics.

7. References

1. *Tasci,Erdal, Tasci,Erdal, Camphausen,Kevin, Krauze,Andra Valentina & Zhuge,Ying. (2022). Glioma Grading Clinical and Mutation Features Dataset. UCI Machine Learning Repository.*
2. *Tasci, Erdal et al. "Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics." International Journal of Molecular Sciences 23 (2022): n. pag.*

8.Acknowldgements

I would like to express my sincere gratitude to Mr. Dhanpratap Sir, my teacher, for his unwavering support and guidance throughout my academic journey. His insightful feedback, constructive criticism, and invaluable advice have been instrumental in shaping my knowledge and skills. I would also like to thank him for inspiring me to push beyond my limits and strive for excellence. Without his encouragement and mentorship, I would not be where I am today. Thank you, Mr. Dhanpratap Sir , for being an exceptional teacher and making a positive impact on my life.