UC Irvine Machine Learning Repository

Glioma Grading Clinical and Mutation Features Dataset

Donated on 12/13/2022

Gliomas are the most common primary tumors of the brain. They can be graded as LGG (Lower-Grade Glioma) or GBM (Glioblastoma Multiforme) depending on the histological/imaging criteria. Clinical and molecular/mutation factors are also very crucial for the grading process. Molecular tests are expensive to help accurately diagnose glioma patients. In this dataset, the most frequently mutated 20 genes and 3 clinical features are considered from TCGA-LGG and TCGA-GBM brain glioma projects. The prediction task is to determine whether a patient is LGG or GBM with a given clinical and molecular/mutation features. The main objective is to find the optimal subset of mutation genes and clinical features for the glioma grading process to improve performance and reduce costs.

Dataset Characteristics
Tabular, Multivariate

Subject Area
Life Science

Associated Tasks
Classification, Other

Attribute Type
Real, Categorical, Integer

 # Instances
839

 # Attributes
23

Information

For what purpose was the dataset created?

Gliomas are the most common primary tumors of the brain. They can be graded as LGG (Lower-Grade Glioma) or GBM (Glioblastoma Multiforme) depending on the histological/imaging criteria. Clinical and molecular/mutation factors are also very crucial for the grading process. Molecular tests are expensive to help accurately diagnose glioma patients. ...

SHOW MORE

Who funded the creation of the dataset?

The Cancer Genome Atlas (TCGA) Project – NCI

What do the instances in this dataset represent?

In this dataset, the most frequently mutated 20 genes and 3 clinical features are considered from TCGA-LGG and TCGA-GBM brain glioma projects.

The preprocessed and organized CSV dataset file consists of twenty-four fields per record. Each field is separated by a comma and each record is separated by a newline. Gender, Age_at_diagnosis, and, Race features are clinical factors, the remaining 20 molecular features consist of IDH1, TP53, ATRX, PTEN, EGFR, CIC, MUC16, PIK3CA, NF1, PIK3R1, FUBP1, RB1, NOTCH1, BCOR, CSMD3, SMARCA4, GRIN2A, IDH2, FAT4, PDGFRA. These molecular features can be mutated or not_mutated (wildtype) depending on the TCGA Case_ID.

Complete attribute documentation for preprocessed dataset file is as follows:

1. Gender : Gender (0 = male; 1 = female)
2. Age_at_diagnosis : Age at diagnosis with the calculated number of days
3. Race : Race
a. 0 = white;
b. 1 = black or african American;
c. 2 = asian;
d. 3 = american indian or alaska native)

4. IDH1 : isocitrate dehydrogenase (NADP(+))1 (0 = NOT_MUTATED; 1= MUTATED)
5. TP53 : tumor protein p53 (0 = NOT_MUTATED; 1 = MUTATED)
6. ATRX : ATRX chromatin remodeler (0 = NOT_MUTATED; 1 = MUTATED)
7. PTEN : phosphatase and tensin homolog (0 = NOT_MUTATED; 1 = MUTATED)
8. EGFR : epidermal growth factor receptor (0 = NOT_MUTATED; 1 = MUTATED)
9. CIC : capicua transcriptional repressor (0 = NOT_MUTATED; 1 = MUTATED)
10. MUC16 : mucin 16, cell surface associated (0 = NOT_MUTATED; 1 = MUTATED)
11. PIK3CA : phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (0 = NOT_MUTATED; 1 = MUTATED)
12. NF1 : neurofibromin 1 (0 = NOT_MUTATED; 1 = MUTATED)
13. PIK3R1 : phosphoinositide-3-kinase regulatory subunit 1 (0 = NOT_MUTATED; 1 = MUTATED)
14. FUBP1 : far upstream element binding protein 1 (0 = NOT_MUTATED; 1 = MUTATED)
15. RB1 : RB transcriptional corepressor 1 (0 = NOT_MUTATED; 1 = MUTATED)
16. NOTCH1 : notch receptor 1 (0 = NOT_MUTATED; 1 = MUTATED)
17. BCOR : BCL6 corepressor (0 = NOT_MUTATED; 1 = MUTATED)
18. CSMD3 : CUB and Sushi multiple domains 3 (0 = NOT_MUTATED; 1 = MUTATED)
19. SMARCA4 : SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 (0 = NOT_MUTATED; 1 = MUTATED)
20. GRIN2A : glutamate ionotropic receptor NMDA type subunit 2A (0 = NOT_MUTATED; 1 = MUTATED)
21. IDH2 : isocitrate dehydrogenase (NADP(+)) 2 (0 = NOT_MUTATED; 1 = MUTATED)
22. FAT4 : FAT atypical cadherin 4 (0 = NOT_MUTATED; 1 = MUTATED)
23. PDGFRA : platelet-derived growth factor receptor alpha (0 = NOT_MUTATED; 1 = MUTATED)

The class label information is given as follows:
- Grade : Glioma grade class information (1 = GBM; 0 = LGG)

Additional Information:

There are 23 instances where Gender, Age_at_diagnosis, or Race feature values are '--', or 'not reported'. These instances, and Project, Case_ID, and Primary_Diagnosis columns were removed from the original dataset file to construct the preprocessed dataset file.

Age_at_diagnosis feature values were converted from string to continuous value by adding day information to the corresponding year information in the dataset as a floating-point number for the preprocessing stage.

All processed and unprocessed files also exist in this directory.

Additional columns of the original dataset file :
Project column represents corresponding TCGA-LGG or TCGA-GBM project names.
Case_ID column refers to the related project Case_ID information.
Primary_Diagnosis column provides information related to the type of primary diagnosis.

SHOW LESS

Are there recommended data splits?
No. We suggest 10-fold cross-validation for feature selection, classification etc.

Does the dataset contain data that might be considered sensitive in any way?
There is race information in this dataset.

Was there any data preprocessing performed?
Yes. There are 23 instances where Gender, Age_at_diagnosis, or Race feature values are '--', or 'not reported'. These instances, and Project, Case_ID, and Primary_Diagnosis columns were removed from the original dataset file to construct the preprocessed dataset file.

Age_at_diagnosis feature values were converted from string to continuous value by adding day information to the corresponding year

information in the dataset as a floating-point number for the preprocessing stage.

All processed and unprocessed files also exist in this directory.

SHOW LESS

Has the dataset been used for any tasks already?

Feature selection, classification etc.

Citation Requests/Acknowledgements

Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., & Krauze, A. V. (2022). Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics. International Journal of Molecular Sciences, 23(22), 14155.

Introductory Paper

Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics

By E. Tasci, Y. Zhuge, Harpreet Kaur, K. Camphausen, A. Krauze. 2022

Published in International Journal of Molecular Sciences

Features

Attribute Name	Role	Type	Description	Units	Missing Values
Grade	Target	Categorical	Grade label	N/A	false
Gender	Feature	Categorical	Gender	N/A	false
Age_at_diagnosis	Feature	Numerical - Continuous	Age at diagnosis with the calculated number of days	years	false
Race	Feature	Categorical	Race (a. 0 = white; b. 1 = black or african American; c. 2 = asian; d. 3 = american indian or alaska native)	N/A	false
IDH1	Feature	Categorical	isocitrate dehydrogenase (NADP(+))1 (0 = NOT_MUTATED; 1= MUTATED)	N/A	false
TP53	Feature	Categorical	tumor protein p53 (0 = NOT_MUTATED; 1 = MUTATED)	N/A	false
ATRX	Feature	Categorical	ATRX chromatin remodeler (0 = NOT_MUTATED; 1 = MUTATED)	N/A	false
PTEN	Feature	Categorical	phosphatase and tensin homolog (0 = NOT_MUTATED; 1 = MUTATED)	N/A	false
EGFR	Feature	Categorical	epidermal growth factor receptor (0 = NOT_MUTATED; 1 = MUTATED)	N/A	false
CIC	Feature	Categorical	capicua transcriptional repressor (0 = NOT_MUTATED; 1 = MUTATED)	N/A	false

Rows per page 10 0 to 10 of 24

DOWNLOAD

CITE

CITE

1 citations
10738 views

Keywords

Brain tumor | Glioma | Mutation | Tumor grading | Clinical features | Molecular features

Creators

Erdal Tasci
erdal.tasci@nih.gov
Radiation Oncology Branch (ROB), National Cancer Institute (NCI), National Institutes of Health (NIH), Building 10

Kevin Camphausen
camphauk@mail.nih.gov
Radiation Oncology Branch (ROB), National Cancer Institute (NCI), National Institutes of Health (NIH), Building 10

Andra Valentina Krauze
andra.krauze@nih.gov
Radiation Oncology Branch (ROB), National Cancer Institute (NCI), National Institutes of Health (NIH), Building 10

Ying Zhuge
zhugey@mail.nih.gov
Radiation Oncology Branch (ROB), National Cancer Institute (NCI), National Institutes of Health (NIH), Building 10

DOI

10.24432/C5R62J

License

Contact

Privacy Notice

Feature Request or Bug Report