scientific reports

Check for updates

OPEN

A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance TCGA data

Raquel Sánchez-Marqués[1,2], Vicente García[3]✉ & J. Salvador Sánchez[4]

Accurate prediction and grading of gliomas play a crucial role in evaluating brain tumor progression, assessing overall prognosis, and treatment planning. In addition to neuroimaging techniques, identifying molecular biomarkers that can guide the diagnosis, prognosis and prediction of the response to therapy has aroused the interest of researchers in their use together with machine learning and deep learning models. Most of the research in this field has been model-centric, meaning it has been based on finding better performing algorithms. However, in practice, improving data quality can result in a better model. This study investigates a data-centric machine learning approach to determine their potential benefits in predicting glioma grades. We report six performance metrics to provide a complete picture of model performance. Experimental results indicate that standardization and oversizing the minority class increase the prediction performance of four popular machine learning models and two classifier ensembles applied on a low-imbalanced data set consisting of clinical factors and molecular biomarkers. The experiments also show that the two classifier ensembles significantly outperform three of the four standard prediction models. Furthermore, we conduct a comprehensive descriptive analysis of the glioma data set to identify relevant statistical characteristics and discover the most informative attributes using four feature ranking algorithms.

Keywords Data-centric machine learning, Glioma grade, Class imbalance, Feature ranking, Clinical factors, Molecular biomarkers

Gliomas are the most common primary tumors of the central nervous system that arise from glial or precursor cells, characterized by increased relapse and mortality rates. Gliomas include astrocytomas, oligodendrogliomas, and ependymomas. According to the 2007 World Health Organization (WHO)[1], astrocytomas are classified into four grades based on the growth potential and aggressiveness. Grades I (pilocytic astrocytomas) and II (diffuse astrocytomas) correspond to the most benign tumors with a favorable prognosis and are considered low-grade gliomas (LGG), whereas grades III (anaplastic astrocytomas) and IV (glioblastomas multiforme, GBM) are considered high-grade gliomas (HGG). Glioblastoma multiforme is the most common, malignant, aggressive, and challenging type of primary brain tumor; it grows rapidly and has the lowest survival rate, with a 5-year survival of around 5%[2]. Since LGG and HGG show different progression and response, and treatment resistance, accurate and early diagnosis and grading are essential to plan appropriate treatment. Furthermore, it should be noted that some subtypes of LGG can lead to GBM in a few months[3], so it is crucial to differentiate LGG from GBM as early as possible.

Currently, the standard procedure for diagnosing, classifying, and grading gliomas is based on histopathological analysis of a sample of brain tissue acquired by surgical biopsy or at the time of resection[4]. However, the potential risks (e.g., the likelihood of damaging a vital brain area can cause neurological deficits) and limitations inherent to biopsy have led to the search for less invasive alternatives without adverse side effects. Thus, significant research efforts have been directed towards the development of neuroimaging techniques that allow the non-invasive extraction of a variety of so-called radiomic features (commonly divided into morphological

[1]Fundación Estatal, Salud, Infancia y Bienestar Social, 28029 Madrid, Spain. [2]Centro de Investigación Biomédica en Red de Enfermedades Infecciosas (CIBERINFEC), Instituto de Salud Carlos III, 28029 Madrid, Spain. [3]Dept. Electrical and Computer Engineering, Instituto de Ingeniería y Tecnología, Universidad Autónoma de Ciudad Juárez, 32310 Ciudad Juárez, Mexico. [4]Dept. Computer Languages and Systems, Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló, Spain. ✉email: vicente.jimenez@uacj.mx

features, textural features, functional features, and semantic features) for the diagnosis, classification of different types of tumors, predicting prognosis and determining the morphology and location of the tumor[5,6]. For instance, Cheng et al.[7] used radiomic features for prediction of glioma grade, Lee et al.[8] for pancreatic cancer, Miranda et al.[9] for rectal cancer, Nguyen et al.[10] for non-small cell lung cancer, Khanfari et al.[11] for prostate cancer grading, and Kim et al.[12] for prediction of disease-free survival in triple-negative breast cancer. In the particular case of glioblastoma, radiomics has emerged as powerful, non-invasive tools to obtain more information about the pathogenesis and therapeutic responses, providing significant biological insights into imaging features[6].

Magnetic resonance imaging (MRI) and computed tomography (CT) are the most commonly used neuroimaging modalities[13]. However, other emerging techniques such as functional MRI (fMRI), magnetic resonance spectroscopy (MRS), positron emission tomography (PET), single-photon emission computed tomography (SPECT), combined PET/CT and hybrid PET/MRI are gaining increasing relevance for the diagnosis, prognosis, and monitoring of gliomas[14–16]. Despite the relevance and usefulness of neuroimaging, advances in genomics and proteomics have allowed the identification of prominent molecular biomarkers that contain both diagnostic and prognostic information for tumors of the central nervous system, becoming a pivotal tool for the evaluation of some gliomas and clinical decision making in neuro-oncology[17]. In 2021, WHO incorporated molecular data as a primary factor in classifying and determining the grade of gliomas, which, together with classic clinical and histological characteristics, can provide better performance[18]. Both methods based on neuroimaging techniques and those that focus on analyzing molecular biomarkers are supported by various machine learning and deep learning models due to their ease in processing large volumes of data and finding the most informative features, as well as their strong performance[19–21].

Deepak and Ameer[22] explored the performance of deep transfer learning with a pre-trained GoogLeNet to extract features from MRI images to discriminate between three types of brain tumors. Analysis of molecular mutations using MRI features proved to be a useful method for diffuse LGG prediction, with the advantage of being a non-invasive procedure[23]. Alksas et al.[24] proposed an imaging-based glioma grading system that uses contrast-enhanced MRI, fluid-attenuated inversion-recovery MRI, and diffusion-weighted MRI to extract morphological, textural, and functional features. Then, the optimal features given by the Gini impurity index are fed to a multi-layer perceptron (MLP) to discriminate between different grades of glioma. Matsiu et al.[25] developed a deep learning model to predict the LGG molecular subtype using a mixture of clinical and radiomic data. An overall accuracy of 68.7% was obtained when the imaging data included MRI, PET, and CT data. Gutta et al.[26] conducted some experiments with a set of 237 patients to demonstrate that the performance of features learned by a convolutional neural network was superior to that of standard radiomic features for glioma grade prediction. Cheng et al.[7] used a total of 2153 intratumoral and peritumoral features extracted from preoperative multiparametric MRI scans of 285 patients to predict glioma grade, reaching an area under the ROC curve (AUC) of 0.975. Furthermore, this technique was shown to have strong generalization performance when applied to an independent validation data set with 65 patients.

Sun et al.[27] compared several radiomic feature selection algorithms and classification models in glioma grading, concluding that the combination of feature selection based on a support vector machine (SVM) with an MLP performed the best in discriminating between LGG and GBM. Cho et al.[28] used the minimum redundancy maximum relevance algorithm with mutual information as the information measure to select the top five features from a total of 468 radiomic features and three classifiers (logistic regression, SVM, and random forest) to distinguish between HGG and LGG images. Bae et al.[29] evaluated the performance and generalizability of traditional machine learning and deep learning models for distinguishing glioblastoma from single brain metastasis using radiomic features. Zhao et al.[30] applied Cox proportional hazards, SVM and random forest to a large glioma data set with 3462 patients for survival prediction, concluding that the best performance was achieved when incorporating radiation therapy and chemotherapy administration status. Tasci et al.[31] introduced a new hierarchical voting-based strategy for feature selection for glioma grading based on clinical and molecular characteristics, improving the performance of using the least absolute shrinkage and selection operator (LASSO) method together with classifier ensembles. Joshi et al.[32] proposed a two-stage ensemble for glioma detection and grading based on clinical and histological data. Munquad et al.[33] employed a correlation-based feature selection scheme and an SVM to predict LGG and subtypes, achieving an average accuracy of 91%. Ren et al.[34] predicted IDH1 (isocitrate dehydrogenase 1) and ATRX (alpha-thalassemia mental retardation X-linked chromatin remodeler) mutations for molecular stratification of LGG using an SVM with a recursive feature elimination algorithm to select an optimal subset of 28 radiomic features. Zheng et al.[35] developed a functional deep neural network to identify high-risk IDH1-mutant glioma patients using clinical factors and molecular features, achieving 90% accuracy.

Zhan et al.[36] proposed a computer-aided diagnosis for grading gliomas which consists of a feature extraction step using PCA to reduce the dimensionality of the data and a prediction step based on a k nearest neighbors classifier. Wu et al.[37] evaluated 50 machine learning algorithms over a data set with 1114 eligible glioma patients and showed that their performance was better than that of the clinical prediction model. The authors concluded that this kind of prediction models can serve as a non-invasive prediction tool for preoperative diagnostic grading of glioma. Ye et al.[38] employed four machine learning algorithms (SVM, random forest, extreme gradient boosting, and generalized linear model) to investigate the relationship between overall survival and the clinical history parameters, pathological characteristics, and molecular alterations of gliomas. The experiments concluded that extreme gradient boosting was the best performing model when applied to a data set with 198 patients. Zhou et al.[39] analyzed the correlation between LGG stemness and clinicopathological characteristics. In addition, the authors used SVM, extreme gradient boosting and LASSO to identify genes critical for stemness subtype prediction. Kha et al.[40] uses Shapley additive explanations (SHAP) analysis to select the best wavelet radiomics features, which were then used with extreme gradient boosting to predict the codeletion status of chromosome 1p/19q in LGG patients.

While most cutting-edge research has focused on the model-building stage of the machine learning process, the performance of a model is highly dependent on data quality. It is now widely accepted that performance improvements are primarily achieved through a data-centric approach[41]. Unlike model-centric systems that focus on how to modify the code, algorithms and representations to improve accuracy and generalization, data-centric approaches focus on curating the data to produce a better performing model. Data-centric machine learning comprises a series of tasks, including standardization and normalization, data cleaning, feature extraction, dimensionality reduction, feature transformation, instance selection, undersampling, data synthesis, and oversampling[42]. However, even recognizing the importance of data-centric methods, the challenge is to find an appropriate balance between these and model-centric methods to provide a robust machine learning solution[43].

This paper aims to present a data-centric approach applied to The Cancer Genome Atlas (TCGA) data set and explore the potential benefits of oversampling and undersampling algorithms to address class imbalance, thus comparing their performance with that of six machine learning models (*k* nearest neighbors, support vector machine, multi-layer perceptron, logistic regression, random forest, and CatBoost). Furthermore, we conduct a comprehensive descriptive analysis of the data set to identify some statistical features and discover the most informative attributes using four feature ranking algorithms (information gain, Gini index, Chi-squared, and random forest). Next, a comparison is carried out with the best performing prediction models using all the features that make up the data set versus the case of using only the five most relevant attributes.

Methodology

This section presents the data set used and its main characteristics, the experimental protocol, and the performance evaluation methods.

Data

All experiments were carried out using a data set[31] obtained from the widely used and publicly available repository of genome atlas data on TCGA (https://www.cancer.gov/tccg). In particular, the data set was built on the basis of the TCGA-LGG and TCGA-GBM projects and consists of three clinical factors (Gender, Age at diagnosis and Race) and 20 frequently mutated molecular biomarkers from 839 patients diagnosed with LGG or GBM. As seen in Table 1, all predictors are categorical type, except for the attribute Age at diagnosis, which is numerical. The molecular features are represented by the values 0 (not mutated) and 1 (mutated) according to the TCGA case number. It is worth noting that it was not necessary to apply any deletion or imputation technique because the data set used in the experiments did not contain missing data on any of the attributes (predictor variables).

#	Predictor	Type	Domain
1	Gender	Clinical	0, 1
2	Age	Clinical	[14.42 . . . 89.29]
3	Race	Clinical	0, 1, 2, 3
4	IDH1	Molecular	0, 1
5	TP53	Molecular	0, 1
6	ATRX	Molecular	0, 1
7	PTEN	Molecular	0, 1
8	EGFR	Molecular	0, 1
9	CIC	Molecular	0, 1
10	MUC16	Molecular	0, 1
11	PIK3CA	Molecular	0, 1
12	NF1	Molecular	0, 1
13	PIK3R1	Molecular	0, 1
14	FUBP1	Molecular	0, 1
15	RB1	Molecular	0, 1
16	NOTCH1	Molecular	0, 1
17	BCOR	Molecular	0, 1
18	CSMD3	Molecular	0, 1
19	SMARCA4	Molecular	0, 1
20	GRIN2A	Molecular	0, 1
21	IDH2	Molecular	0, 1
22	FAT4	Molecular	0, 1
23	PDGFRA	Molecular	0, 1

Table 1. Information on the 23 predictors in the data set.

Descriptive statistics

The data set consists of two classes indicating the glioma grade: 487 (58.05%) patients with LGG (the positive class, 0) and 352 (41.95%) with GBM (the negative class, 1), resulting in an imbalance ratio of 1.38 (i.e., the ratio of majority to minority samples in the data set). Of the total samples in the data set, 488 (58.16%) correspond to men (0) and 351 (41.84%) to women (1). Regarding the attribute Race, there are 765 cases of white people (0), 59 of black or African American people (1), 14 of Asians (2), and only 1 American Indian (3). Table 2 reports the distribution of cases according to glioma grades for the clinical factors. The mean age values in Table 2 suggest that there are no significant differences between males and females affected by these brain tumors, even regardless of the glioma grade. On the other hand, the data regarding patient race could be biased because the vast majority of cases are white people, so any conclusions about the incidence of glioma based on the attribute Race could be erroneous.

Table 3 summarizes a series of descriptive statistics for the attribute Age at diagnosis according to the gender of the patients, including measures of central tendency and measures of dispersion: minimum and maximum values, arithmetic mean, median, mid range, standard deviation (SD), standard error (SE), 95% confidence interval (95% CI), first (Q1) and third (Q3) quartiles, interquartile range (IQR), coefficient of skewness, coefficient of kurtosis, kurtosis excess, and coefficient of variation (CV). Additionally, we conducted Kolmogorov-Smirnov (K-S) test[46] (with Lilliefors significance correction) at a significance level of 0.05 to check for normality of the distribution of the samples in each gender; if p-value > 0.05, it may be assumed that the data follow a normal distribution. We chose the K-S test instead of the Shapiro–Wilk test because it is more appropriate for large sample size ($N \geq 50$)[47].

To visualize the shape of the distributions, Fig. 1 shows histograms and density plots for the attribute Age at diagnosis for both males and females. In addition, it also displays the Q-Q plots for the attribute Age at diagnosis.

	Total (N = 839)	LGG (N = 487)	GBM (N = 352)
Gender			
Male	488 (58.16)	271 (55.65)	217 (61.65)
Age at diagnosis	51.15 (15.81)	43.32 (13.52)	60.94 (12.72)
Female	351 (41.84)	216 (44.35)	135 (38.35)
Age at diagnosis	50.63 (15.57)	44.57 (12.92)	60.33 (14.53)
Race			
White	765 (91.18)	457 (93.84)	308 (87.50)
Black/African American	59 (7.03)	21 (4.31)	38 (10.80)
Asian	14 (1.67)	8 (1.64)	6 (1.70)
American Indian	1 (0.12)	1 (0.21)	0 (0.00)

Table 2. Distribution of cases (N (%)) according to glioma grades based on clinical factors. For Age at diagnosis, the mean and (standard deviation) are shown.

	Male (N = 488)	Female (N = 351)
Lowest value	14.42	20.32
Highest value	89.29	85.61
Mean	51.1528	50.6331
Median	52.1250	50.3500
Mid range	51.86	52.97
SD	15.81075	15.56781
SE	0.03240	0.04436
95% CI Lower bound	49.7466	48.9988
Upper bound	52.5591	52.2674
Q1	38.14	38.04
Q3	63.24	62.11
IQR	25.14	24.07
Coefficient of skewness	0.04337	0.11670
Coefficient of kurtosis	2.20588	2.17555
Kurtosis excess	− 0.81268	− 0.85033
CV	0.30909	0.30746
K-S statistic	0.0615	0.0545
K-S p-value	0.0496	0.5801

Table 3. Descriptive statistics of the attribute Age at diagnosis.

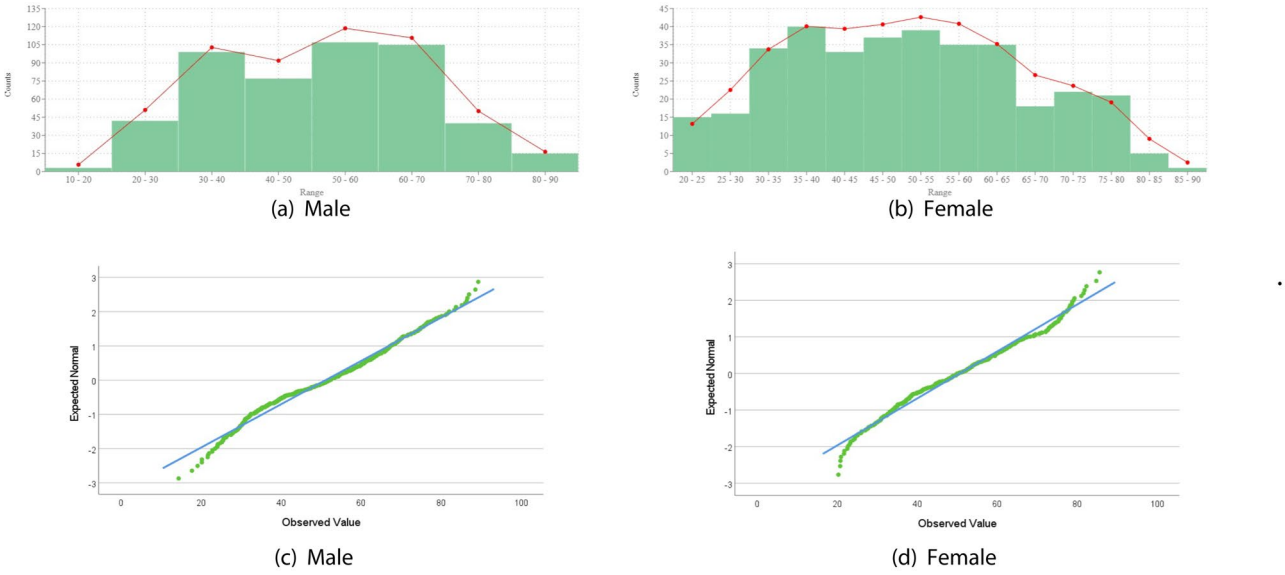

Figure 1. Histograms (green boxes), density plots (red line) and Q-Q plots for the attribute Age at diagnosis.

As can be seen, the residuals (green dots) tend to deviate quite a bit from the 45-degree line (blue) only at the tail ends, indicating that the data follow a normal distribution.

On the other hand, Fig. 2 shows a box-plot with the distribution of the attribute Age at diagnosis for LGG (class 0) and GBM (class 1) cases. The dark blue vertical and the thin blue lines represent the mean age and standard deviation, respectively. The median is shown with a yellow vertical line, while the blue highlighted area represents the values between the first and the third quartiles.

Table 4 displays counts (frequencies) and proportions (relative frequencies) for the 20 molecular biomarkers. Note that the list is ordered from highest to lowest by the count (or percentage) of cases with a mutation in the corresponding biomarker, ranging from 404 for IDH1 to 22 for PDGFRA.

Experimental protocol

The multiple machine learning models used to predict glioma grade in the experiments included four standard classification models and two powerful classifier ensembles. The standard models were *k*-nearest neighbors (kNN), SVM, MLP, and logistic regression (LR). The ensembles were random forest (RF) and CatBoost. kNN is a non-parametric learning algorithm that produces the class label of an input sample based on the majority vote of its *k* closest training cases. SVM is a supervised machine learning model that classifies data by finding the hyperplane that optimally separates the samples of one class from the other, that is, the hyperplane that maximizes the distance (margin) between the closest samples of the opposite class. One of the most interesting features of SVM is that it works for both linear and nonlinear problems, as well as being less prone to overfitting. When data are not linearly separable, some kernel function must be used to transform the training data into a higher-dimensional feature space that allows linear separability. An MLP is an artificial neural network that consists of multiple layers of interconnected neurons: an input layer that receives the input sample as a combination of the feature values, an output layer that performs the classification by using some activation function, and one

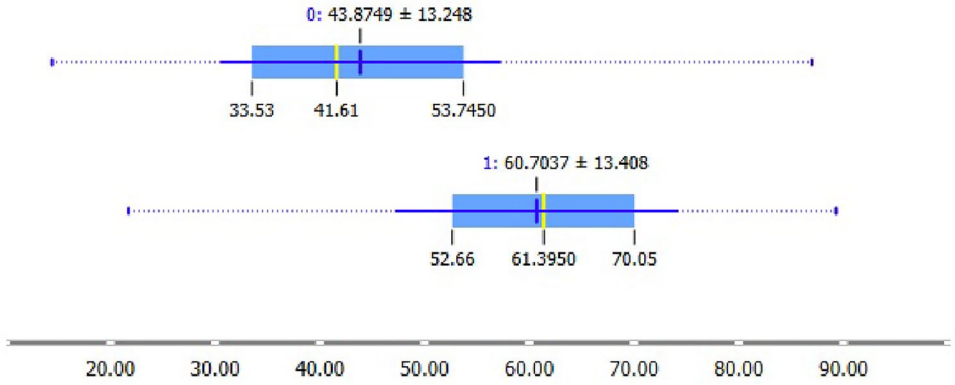

Figure 2. Distribution of values for the attribute Age at diagnosis.

Biomarker	Non-mutated (0)		Mutated (1)	
	Count	Proportion (%)	Count	Proportion (%)
IDH1	435	51.85	404	48.15
TP53	491	58.52	348	41.48
ATRX	622	74.14	217	25.86
PTEN	698	83.19	141	16.81
EGFR	727	86.65	112	13.35
CIC	728	86.77	111	13.23
MUC16	741	88.32	98	11.68
PIK3CA	766	91.30	73	8.70
NF1	772	92.01	67	7.99
PIK3R1	785	93.56	54	6.44
FUBP1	794	94.64	45	5.36
RB1	799	95.23	40	4.77
NOTCH1	801	95.47	38	4.53
BCOR	810	96.54	29	3.46
CSMD3	812	96.78	27	3.22
SMARCA4	812	96.78	27	3.22
GRIN2A	812	96.78	27	3.22
IDH2	816	97.26	23	2.74
FAT4	816	97.26	23	2.74
PDGFRA	817	97.38	22	2.62

Table 4. Frequencies and proportions for the molecular biomarkers.

or more hidden layers (placed in between the input and output layers) whose neurons perform computations on the inputs. The logistic regression model makes a prediction based on the probability that an input sample belongs to a particular class: if the probability is greater than 0.5, the sample is assigned to that class; otherwise, the sample is classified to the other class.

RF[44] is an extension of the bagging method made up of multiple decision trees, each generated from a sample drawn with replacement from the training set (i.e., with replacement means that one sample could be selected multiple times, while others could not be selected at all). During the construction of a tree, the best split is selected from a random subset of features, thus ensuring low correlation between decision trees. When classifying new input samples, all trees make a judgment and the final decision is made by majority vote. CatBoost[45] is an improved implementation of gradient boosting on binary decision trees, which means that each new tree is trained to minimize the loss function of the previous model (i.e., to reduce the error made by previous trees) using gradient descent. CatBoost handles categorical features not by using a binary substitution of the categorical values but by performing a random permutation of the training data (this ensures different orderings during different stages of the gradient boosting process) and calculating the average label value for the sample with the same class value placed before the given one in the permutation.

Before applying the prediction models, the values of the attribute Age at diagnosis were normalized using the z-score standardization technique so that the mean of all values was 0 and the standard deviation was 1. A raw value x of the feature is converted into a normalized value z by

$$z = \frac{x - \bar{x}}{SD} \tag{1}$$

where $\bar{x}$ and SD are the mean and standard deviation of a feature, respectively.

Note that normalization was applied solely to Age at diagnosis because all other attributes were categorical. On the other hand, to find the best values of the hyperparameters for the machine learning models, we fine-tuned them using an 80-20 stratified holdout setting method (Table 5).

Performance evaluation

We adopted a stratified 10-fold cross-validation method, where the data set was randomly divided into ten stratified non-overlapping blocks of roughly equal size. The models were trained with nine of these blocks combined and then applied to the remaining block to estimate the performance. This process was repeated for each of the 10 blocks, giving a total of 755 training samples and 84 testing samples in each of the 10 iterations of the cross-validation. Performance was then calculated as the average of the 10 estimates thus obtained. We used six scalar indicators to evaluate the prediction performance: classification accuracy (Acc), Precision (Prec), Recall, Specificity (Spec), F1-score (F1), and Matthews correlation coefficient (MCC). All these measures were derived from a 2×2 confusion matrix, where each entry (i, j) represents the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Model	Parameter	Settings	Best value	Model	Parameter	Settings	Best value
kNN	Neighbors (*k*)	[1 . . . 15]	5	RF	Number of trees	[80, 100, 120]	120
	Distance metric		Euclidean		Number of attributes considered at each split	[1 . . . 23]	11
SVM	Cost (C)	[0.01 . . . 1]	1		Limit depth of trees	[3 . . . 10]	5
	Kernel	Linear, RBF	Linear	CatBoost	Estimator	Tree	Tree
MLP	Hidden layers	[1 . . . 3]	1		Number of trees	[80, 100, 120]	100
	Neurons in hidden layer	[50, 100, 150]	50		Learning rate	[0 . . . 1]	0.03
	Activation function	Logistic, ReLu	ReLu		Limit depth of trees	[3 . . . 10]	6
	Learning iterations	[50 . . . 200]	100		Regularization type	LASSO, Ridge	Ridge
LR	Regularization type	LASSO, Ridge	Ridge				

Table 5. Model hyperparameters.

In addition to these scalar metrics, we also included the receiver operating characteristics (ROC) and the precision-recall curves to visualize how the machine learning models used in the experiments perform in predicting the classes. The ROC curve plots a false positive rate (i.e., 1-specificity) on an X-axis against a true positive rate on a Y-axis; the closer the curve approaches the upper left corner of the ROC space, the better the model is at predicting the classes. The precision-recall curve shows the ratio between precision (ratio of true positives in positive predictions) and recall (ratio of true positives in positive class) at different thresholds; ideally, the curve should be as close to the top right corner as possible.

Results and discussion

This section consists of three blocks. First, we investigated the most informative molecular biomarkers based on the distribution of cases in each glioma grade and checked whether these results agree with the results of four feature ranking algorithms. The second block analyzes the performance of six standard prediction models and classifier ensembles for glioma grading. Finally, we apply some resampling techniques to handle class imbalance and verify if this leads to increased performance.

Most informative features

Values in Table 6 show the distribution of cases according to glioma grades for the mutated molecular biomarkers (i.e., feature value = 1). As can be seen, IDH1 mutations are the most common, being detected in 404 patients

Predictor	Total (N = 839)	LGG (N = 487)	GBM (N = 352)
IDH1	**404 (48.15)**	**381 (94.31)**	**23 (5.69)**
TP53	348 (41.48)	235 (67.53)	113 (32.47)
ATRX	**217 (25.86)**	**183 (84.33)**	**34 (15.67)**
PTEN	**141 (16.81)**	**25 (17.73)**	**116 (82.27)**
EGFR	112 (13.35)	31 (27.68)	81 (72.32)
CIC	**111 (13.23)**	**107 (96.40)**	**4 (3.60)**
MUC16	98 (11.68)	41 (41.84)	57 (58.16)
PIK3CA	73 (8.70)	39 (53.42)	34 (46.58)
NF1	67 (7.99)	29 (43.29)	38 (56.72)
PIK3R1	54 (6.44)	21 (38.89)	33 (61.11)
FUBP1	45 (5.36)	43 (95.56)	2 (4.44)
RB1	40 (4.77)	6 (15.00)	34 (85.00)
NOTCH1	38 (4.53)	38 (100)	0 (0.00)
BCOR	29 (3.46)	17 (58.62)	12 (41.38)
CSMD3	27 (3.22)	12 (44.44)	15 (55.56)
SMARCA4	27 (3.22)	23 (85.19)	4 (14.81)
GRIN2A	27 (3.22)	7 (25.93)	20 (74.07)
IDH2	23 (2.74)	21 (91.30)	2 (8.70)
FAT4	23 (2.74)	11 (47.83)	12 (52.17)
PDGFRA	22 (2.62)	6 (27.27)	16 (72.73)

Table 6. Distribution of cases (N (%)) according to glioma grades based on mutated molecular biomarkers (bold values indicate the most discriminating molecular biomarkers, that is, those with the greatest difference between LGG cases and GBM cases).

(48.15% of the total cases studied). However, these mutations occur in 94.31% of cases with LGG and only in 5.69% of cases with GBM, confirming previous findings that this is a very informative molecular biomarker for glioma grading[17,34,48]: IDH1/2 mutations have been largely associated with grade II and III gliomas and secondary glioblastomas[49]. Looking at the biomarkers with 50 or more cases, similar conclusions can be drawn for the molecular biomarkers ATRX with 84.33% of LGG, PTEN (phosphatase and tensin homolog) with 82.27% of patients affected by GBM and CIC (capicua transcriptional repressor) with 96.40% LGG. In the case of biomarkers with a low percentage of patients, we find NOTCH1 (notch receptor 1) (100% of LGG), FUBP1 (far upstream element binding protein 1) (95.56% of LGG), IDH2 (isocitrate dehydrogenase 2) (91.30% of LGG), SMARCA4 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4) (85.19% of LGG) and RB1 (retinoblastoma transcriptional corepressor 1) (85.00% of GBM).

To support the conclusions drawn from the values in Table 6, we ran four feature ranking algorithms on the normalized data set with the aim of checking which are the most informative predictors: information gain, Gini index, Chi-squared, and RF. Note that identifying the most informative clinical factors and glioma molecular biomarkers can be valuable in obtaining relevant biological information. On the other hand, in some practical cases, having small feature sets with high prediction accuracy can become paramount to minimize response time.

Information gain (infGain) estimates the relevance of a predictor based on the amount by which the entropy of the class decreases when considering that feature. The Gini index (Gini) estimates the distribution of a predictor in different classes and can be interpreted as a measure of impurity for a feature. Chi-squared (Chi2) measures the relationship strength between each variable and the class label. Note that Chi-squared applies to categorical predictors, and therefore, numerical attributes (as is the case for Age at diagnosis) must first be discretized into several intervals. In the case of RF as a feature ranking method, each tree in the forest calculates the importance of a predictor based on its ability to decrease the weighted impurity in the tree.

Since each feature ranking algorithm could yield different results (rankings), fusing them using a multiple intersection method was necessary to find out which features got the highest rankings in the output of the four algorithms. Thus, looking at the rankings of each algorithm, it was possible to determine which were the five most relevant attributes, while there were discrepancies in establishing the most informative attributes from the sixth position onwards. From the outputs of the multiple intersection method, Table 7 shows that the four feature ranking algorithms agreed to define IDH1 as the most informative attribute, followed by Age at diagnosis, PTEN, CIC and ATRX. These results are interesting because they are consistent with the findings of various studies conducted in neuroscience and neuro-oncology[17,34,48,49] in which the mutated molecular biomarkers that best discriminate LGG from GBM were determined, as reported in Table 6. The relevance of this lies in the fact that feature selection or ranking algorithms could be used to discover molecular biomarkers with the greatest discriminating power instead of other more expensive, time-consuming and difficult to carry out methods.

Predictor	infGain	Gini	Chi2	RF
Gender	0.003 (19)	0.002 (19)	1.759 (20)	0.010 (13)
Age	**0.209 (2)**	**0.130 (2)**	**185.221 (2)**	**0.201 (2)**
Race	0.012 (11)	0.008 (11)	6.836 (17)	0.011 (10)
IDH1	**0.414 (1)**	**0.244 (1)**	**218.137 (1)**	**0.245 (1)**
TP53	0.019 (10)	0.013 (10)	12.852 (10)	0.017 (7)
ATRX	**0.078 (5)**	**0.048 (4)**	**61.571 (5)**	**0.029 (4)**
PTEN	**0.100 (3)**	**0.066 (3)**	**94.102 (3)**	**0.025 (5)**
EGFR	0.042 (6)	0.028 (6)	42.410 (6)	0.016 (8)
CIC	**0.085 (4)**	**0.045 (5)**	**67.040 (4)**	**0.040 (3)**
MUC16	0.010 (14)	0.007 (12)	10.572 (12)	0.008 (17)
PIK3CA	0.001 (22)	0.000 (22)	0.640 (22)	0.010 (12)
NF1	0.006 (18)	0.004 (18)	5.995 (18)	0.011 (11)
PIK3R1	0.007 (17)	0.005 (16)	8.137 (16)	0.006 (18)
FUBP1	0.030 (8)	0.016 (9)	26.000 (9)	0.008 (15)
RB1	0.029 (9)	0.019 (7)	30.434 (7)	0.013 (9)
NOTCH1	0.037 (7)	0.017 (8)	27.466 (8)	0.008 (14)
BCOR	0.000 (23)	0.000 (23)	0.004 (23)	0.008 (16)
CSMD3	0.002 (20)	0.001 (20)	2.051 (19)	0.005 (19)
SMARCA4	0.008 (15)	0.005 (17)	8.166 (15)	0.001 (23)
GRIN2A	0.010 (13)	0.007 (13)	11.438 (11)	0.005 (20)
IDH2	0.011 (12)	0.006 (14)	10.447 (13)	0.023 (6)
FAT4	0.001 (21)	0.001 (21)	0.986 (21)	0.003 (22)
PDGFRA	0.008 (16)	0.005 (15)	8.555 (14)	0.005 (21)

Table 7. Results of feature ranking methods. The ranking of each feature is shown in brackets (bold values indicate the five most informative predictors based on the multiple intersection method).

We ran multidimensional scaling[50] to visualize in Fig. 3 the samples from both classes as a function of the attribute Age at diagnosis against each of the four most informative biomarkers (IDH1, PTEN, CIC, and ATRX). Each blue dot represents an LGG sample, and each red dot is a GBM sample. The regions belonging to each class are shaded in blue or red depending on whether they correspond to the LGG or GBM class, respectively. These graphs allow us to see how the age of the patients and mutations are related to the grade of glioma. For example, Fig. 3a reveals that most LGG cases require IDH1 mutations and occur at younger ages than GBM cases. For PTEN (Fig. 3b), LGG occurs when there is no mutation, while GBM does not appear to depend on this molecular biomarker since approximately the same number of cases is seen both with and without PTEN mutations.

Results of the prediction models

Table 8 reports the results of each of the six evaluation metrics achieved by the prediction models applied to the normalized data set (with all predictors) using the experimental protocol described above. The results revealed that RF was the best performing model, although closely followed by CatBoost and SVM. In contrast, kNN. MLP and LR obtained the lowest values regardless of the performance evaluation metric used.

For better analysis of these results, we performed a pairwise comparison of models using a correlated Bayesian *t*-test[51] for each evaluation metric to check whether the difference in scores between each pair of models was significant or not. Unlike the frequentist correlated *t*-test, where the inference is a *p*-value, the inference of the Bayesian *t*-test is a posterior probability. Additionally, this test considers the correlation and the uncertainty (i.e., the standard error) of the results generated by cross-validation. The outputs of the statistical test are summarized in Table 9, where the number in a cell denotes the probability that the model corresponding to the row had a significantly higher score (posterior probability greater than 0.5) than the model corresponding to the column. Values in this table indicate that the results obtained by RF and CatBoost were significantly better than those of kNN, MLP, and LR, regardless of the metric used. When comparing RF and CatBoost with SVM, it can be seen that the differences were not statistically significant when using Prec (0.492 and 0.399) and Spec (0.460 and 0.416). Finally, posterior probabilities of RF being significantly better than CatBoost revealed that the performance differences between both ensembles were very small, so one should not conclude that RF performed better than CatBoost.

Figure 4 plots the ROC curves for the RF and CatBoost ensembles separately for each of the two classes (LGG and GBM). The diagonal dotted line represents the behavior of a random classifier, while the full diagonal line represents iso-performance in the ROC space so that all the points on the line give the same profit/loss. The closer to the top and further to the left this full diagonal line is, the better the classifier result. The AUC was 0.923 for RF and 0.924 for CatBoost, that is, the difference between both classifiers was negligible.

Figure 5 shows the confusion matrix corresponding to each of the six prediction models. Although it was seen that the imbalance ratio of the data set was moderately low (1.38), the confusion matrix allows us to discover the behavior of the models in each of the classes, that is, analyze the number of successes and errors individually by class in order to identify whether or not there were differences between predicting samples belonging to the majority class and samples of the minority class. Thus, it can be observed that the three models with the best performance (RF, CatBoost and SVM) made a lower number of errors than the other three classifiers (kNN,

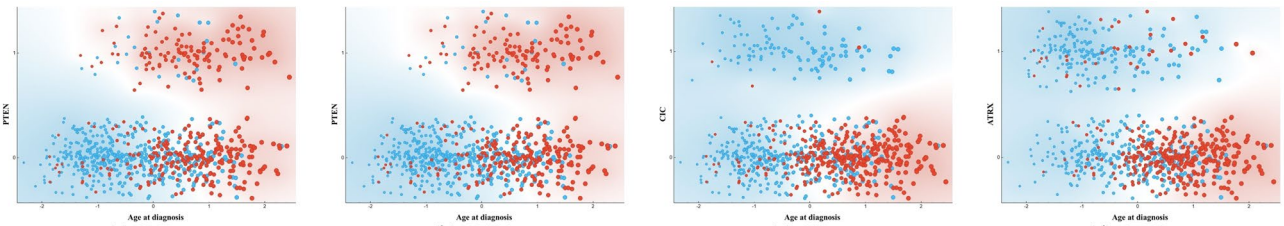

Figure 3. Scatter plot of Age at diagnosis (X-axis) vs. the most informative molecular biomarkers (Y-axis).

Model	Acc	F1	Prec	Recall	MCC	Spec
kNN	0.852	0.853	0.856	0.852	0.702	0.856
SVM	0.867	0.867	**0.878**	0.867	0.741	**0.884**
MLP	0.852	0.853	0.857	0.852	0.703	0.859
LR	0.862	0.862	0.865	0.862	0.721	0.866
RF	**0.869**	**0.870**	**0.878**	**0.869**	**0.743**	0.883
CatBoost	0.869	**0.870**	0.877	**0.869**	0.742	0.882
Mean	0.862	0.863	0.869	0.862	0.725	0.872
SD	0.008	0.008	0.011	0.008	0.019	0.013

Table 8. Prediction performance of the machine learning models (the best values are in bold).

	kNN	SVM	MLP	LR	RF	CatBoost		kNN	SVM	MLP	LR	RF	CatBoost
Accuracy							Recall						
kNN		0.093	0.500	0.286	0.020	0.072	kNN		0.093	0.500	0.286	0.020	0.072
SVM	0.907		0.895	0.643	0.354	0.324	SVM	0.907		0.895	0.643	0.354	0.324
MLP	0.500	0.105		0.236	0.077	0.037	MLP	0.500	0.105		0.236	0.077	0.037
LR	0.714	0.357	0.764		0.306	0.281	LR	0.714	0.357	0.764		0.306	0.281
RF	0.980	0.646	0.923	0.694		0.501	RF	0.980	0.646	0.923	0.694		0.501
CatBoost	0.928	0.676	0.963	0.719	0.499		CatBoost	0.928	0.676	0.963	0.719	0.499	
F1-score							Matthews correlation coefficient						
kNN		0.091	0.493	0.284	0.020	0.070	kNN		0.039	0.504	0.286	0.013	0.056
SVM	0.909		0.894	0.647	0.354	0.320	SVM	0.961		0.958	0.790	0.440	0.481
MLP	0.507	0.106		0.240	0.078	0.037	MLP	0.496	0.042		0.242	0.046	0.019
LR	0.716	0.353	0.760		0.302	0.275	LR	0.714	0.210	0.758		0.219	0.203
RF	0.980	0.646	0.922	0.698		0.497	RF	0.987	0.560	0.954	0.781		0.544
CatBoost	0.930	0.680	0.963	0.725	0.503		CatBoost	0.944	0.519	0.981	0.797	0.456	
Precision							Specificity						
kNN		0.019	0.515	0.290	0.010	0.049	kNN		0.013	0.437	0.281	0.009	0.033
SVM	0.981		0.980	0.876	0.508	0.601	SVM	0.987		0.976	0.907	0.540	0.584
MLP	0.485	0.020		0.241	0.031	0.012	MLP	0.563	0.024		0.303	0.037	0.013
LR	0.710	0.124	0.759		0.167	0.157	LR	0.719	0.093	0.697		0.128	0.115
RF	0.990	0.492	0.969	0.833		0.579	RF	0.991	0.460	0.963	0.872		0.536
CatBoost	0.951	0.399	0.988	0.843	0.421		CatBoost	0.967	0.416	0.987	0.885	0.464	

Table 9. Pairwise comparison of models.

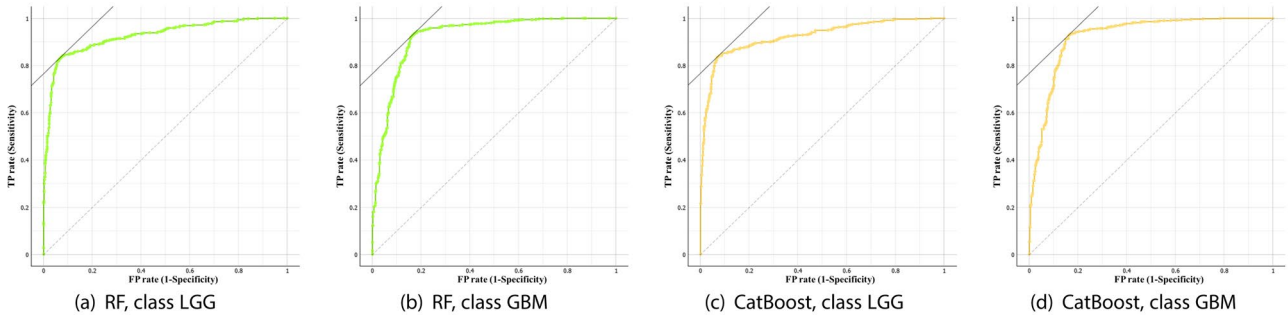

Figure 4. ROC curves for the classifier ensembles.

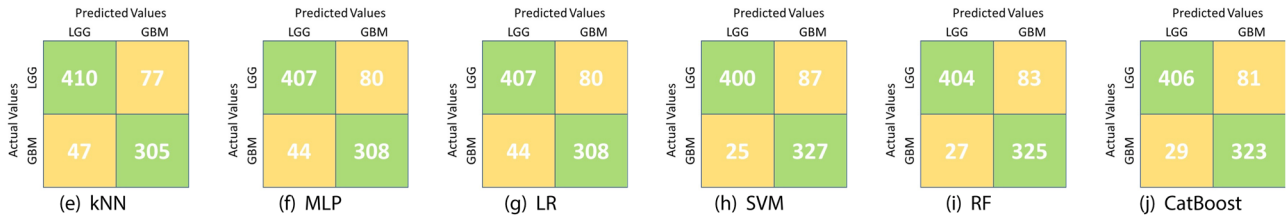

Figure 5. Confusion matrices of the classifiers.

MLP and LG) on the minority class (GBM). In contrast, the number of misclassifications on the majority class (LGG) was similar in all classifiers.

Explainability of predictions

Due to the "black box" nature of most machine learning models, one of the main problems is their insufficient interpretability or the difficulty in understanding the predictions they make. To shed light on these limitations, some methodologies belonging to the eXplainable Artificial Intelligence (XAI)[52] paradigm have been proposed in order to provide a reasonable understanding of the output of machine learning models. In particular, we

analyzed the effect of the attributes on the prediction performance using two explainability approaches: global feature importance and SHAP.

Global feature importance estimates the contribution of each individual feature to the prediction by measuring the increase in the prediction error of the model after performing permutations on the feature values across the data set, which breaks the relationship between the feature and the target variable[44,53]. A feature is important if permuting its values increases the model error, while a feature is of little or no importance if permuting its values does not change the error of the model.

Bar charts in Fig. 6 show feature importances in descending order for each classifier, indicating that the IDH1 biomarker was the most important attribute contributing to the target variable (i.e., glioma grade), regardless of the model used. The second most important feature was Age at diagnosis in all cases except when applying the MLP neural network (note that even in this case the attribute Age at diagnosis was the third most important). It is worth highlighting that these results mostly agree with those reported in Table 7, where these two features were also identified as the most relevant when applying the multiple intersection method.

It should be noted that the global feature importance approach reveals the absolute importance of each attribute, but it does not indicate the direction of the change given by the permutations, that is, it does not report whether the feature increases or decreases the prediction performance of the model. To overcome this limitation, we also employed the SHAP method introduced by Lundberg and Lee[54], which is based on the principles of cooperative game theory and can provide broad explanations of model predictions at both local and global levels. This method computes Shapley values, which quantify the average marginal contribution of a feature to the prediction made by the model after considering all possible combinations with other features[55], that is, it provides information about whether the influence of each characteristic on the prediction value of the model is positive (increase) or negative (decrease). The Shapley value of a feature, is calculated as the difference between the prediction when the feature is present and the prediction when the feature is absent.

Figure 7 shows the SHAP summary plot for each model, which represents the positive or negative impact of each feature on the prediction of one class. On the X-axis is the Shapley value, which denotes how much the features contribute to the prediction of a patient diagnosed with GBM across all possible combinations. A value less than 0 indicates a negative contribution (i.e., low importance for the prediction of the minority class GBM), equal to 0 indicates no contribution, and greater than 0 indicates a positive contribution (i.e., high importance for prediction). The left vertical axis (Y-axis) is for features ranked in descending order of their relevance to the prediction of class GBM, while the right vertical axis indicates the value of the features from lowest to highest. Each dot represents the Shapley value of a sample (patient) plotted horizontally and is colored red or blue depending on whether the value is high or low, respectively.

From these plots, it can be seen that Age at diagnosis was the most important feature for the prediction in class GBM when the KNN and LR models were used, and the second most relevant with the rest of the classifiers. Samples with higher values of this feature (red color) had higher Shapley values, meaning that they contributed to the prediction of class GBM. Lower values of this attribute (blue) contributed against the prediction of this class. The IDH1 biomarker (categorical attribute) contributed the most to the prediction of GBM class when using the MLP, SVM, RF and CatBoost models. As IDH1 is a categorical attribute, its impact on the prediction depends on its value (0 = non-mutated, 1 = mutated). Thus, it can be seen that this biomarker with the non-mutated value for the patient (red color) contributed to the prediction of the GBM class, while the mutated value of this attribute contributed negatively.

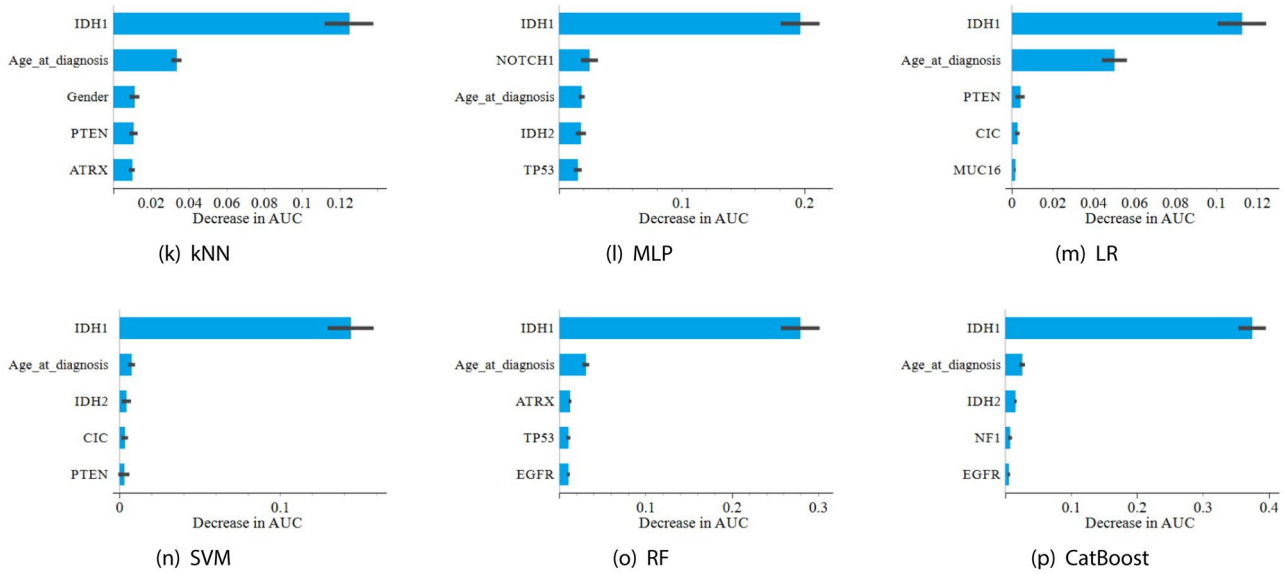

Figure 6. Feature importance of the top 5 variables according to the AUC of the model.

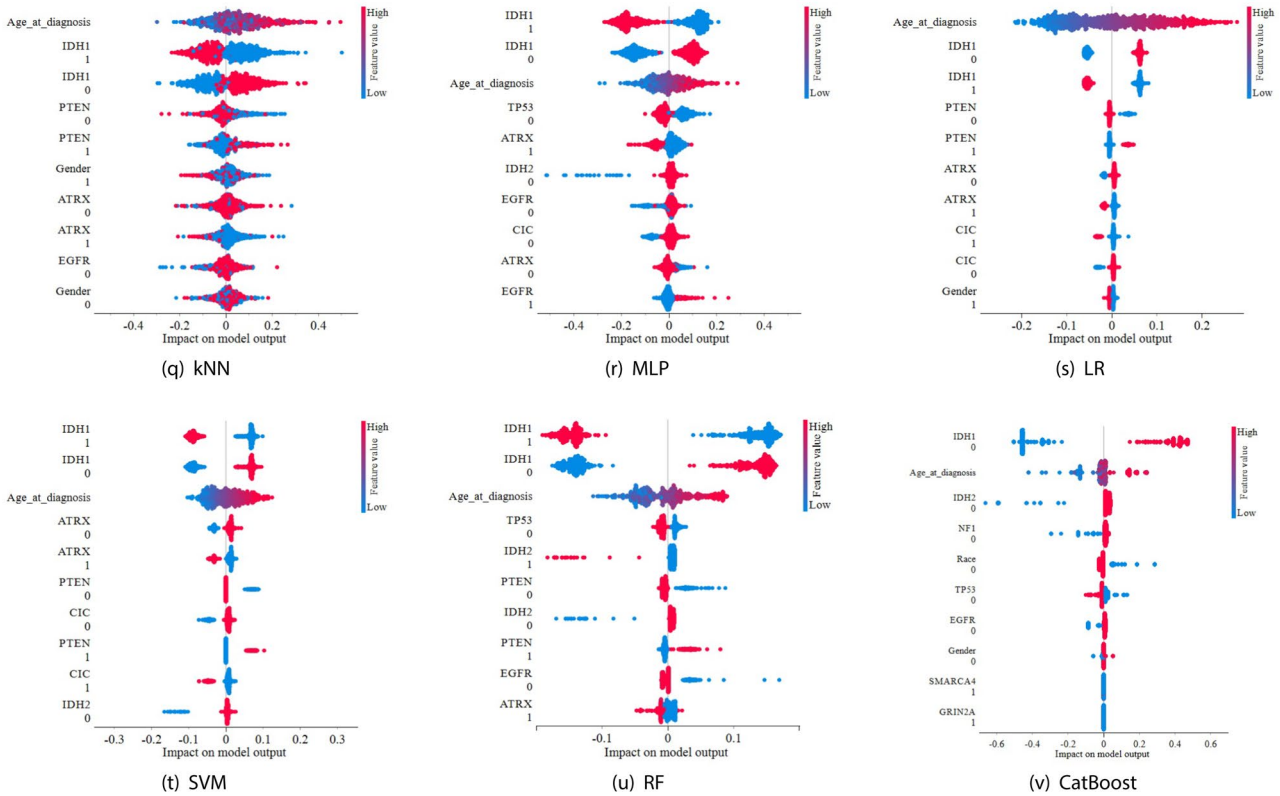

Figure 7. SHAP summary plots.

Addressing class imbalance

Considering the differences in misclassifications between the majority class and the minority class, we decided to address the class imbalance in order to see if any performance improvement could be obtained. It is well known that training a machine learning algorithm with imbalanced data can favor the majority class, typically leading to higher misclassification rates over the minority class (GBM). Among the various strategies to address imbalanced data, resampling techniques are by far the most widely used approach because they have been proven to be efficient, classifier-independent, and can be easily implemented for any problem[56]. These are designed to change the composition of the training data set by adjusting the number of majority and/or minority samples until both classes are represented by an approximately equal number of samples. Many researchers have argued that over-sampling is generally superior to under-sampling because under-sampling algorithms can discard potentially useful data and increase classifier variance[57]. It should be noted that, to avoid overoptimistic results, resampling should be applied only to the training set, not to the entire data set[58]. In the case of over-sampling, for instance, this means that the testing samples are neither over-sampled nor seen by the machine learning model during training.

Experiments in this section were carried out with two resampling algorithms. The first is an over-sampling algorithm proposed by Chawla et al.[59] called SMOTE, which generates artificial samples of the minority class (GBM) by interpolating existing samples that are close together. It first finds the k minority nearest neighbors for each minority sample, and then synthetic samples are generated in the direction of some or all of those nearest neighbors. Depending on the amount of over-sampling required, a certain number of samples are randomly chosen from the k nearest neighbors. The second is random under-sampling (RUS), which balances the data set by randomly removing samples that belong to the over-sized class (LGG).

Table 10 reports the performance results obtained after preprocessing the normalized data set with SMOTE and RUS. The first issue worth mentioning is that oversampling performed better than undersampling, except when Recall was used. Secondly, unlike the results obtained with the normalized data set without preprocessing (Table 8), now the best model after up-sampling the data set was SVM, although the differences concerning RF and CatBoost were really negligible.

To check whether or not the difference in the means of the results with the normalized training set without preprocessing and those preprocessed with over-sampling and under-sampling were significant, a two-tailed t-test[60] was performed for a significance level of 5% ($\alpha = 0.05$), whose t-values and p-values are shown in Table 11. Thus, when comparing the means of Table 8 with those of over-sampling (upper part of Table 10), we obtained that the differences were statistically significant in all cases, except when using the specificity to evaluate the performance of the models. On the other hand, when comparing them with the means obtained with under-sampling (bottom of Table 10), we found that the differences in precision and specificity on the non-preprocessed set were significantly better than those of the downsized set. Therefore, despite the low imbalance index, the test

Over-sampling (SMOTE)							Under-sampling (RUS)						
Model	**Acc**	**F1**	**Prec**	**Recall**	**MCC**	**Spec**	**Model**	**Acc**	**F1**	**Prec**	**Recall**	**MCC**	**Spec**
kNN	0.868	0.867	0.870	0.868	0.738	0.868	kNN	0.854	0.859	0.828	0.892	0.709	0.815
SVM	**0.882**	**0.882**	**0.886**	**0.882**	**0.768**	**0.882**	SVM	0.879	0.885	0.845	**0.929**	0.762	0.830
MLP	0.862	0.862	0.864	0.862	0.727	0.862	MLP	0.851	0.855	0.831	0.881	0.703	0.821
LR	0.871	0.870	0.873	0.871	0.744	0.871	LR	0.866	0.872	0.838	0.909	0.736	0.824
RF	0.881	0.881	0.885	0.881	0.766	0.881	RF	0.876	0.883	0.841	**0.929**	0.757	0.824
CatBoost	0.881	0.881	0.885	0.881	0.766	0.881	CatBoost	**0.881**	**0.886**	**0.847**	**0.929**	**0.765**	**0.832**

Table 10. Prediction performance of the machine learning models using the resampled data sets (the best values are in bold).

Method	Acc	F1	Prec	Recall	MCC	Spec
SMOTE	11.055865	10.159443	7.985837	11.055865	12.726012	*1.119902*
	< .01	< .01	< .01	< .01	< .01	*0.31365*
RUS	2.757831	4.364481	−17.731469	9.101624	3.779645	−13.991676
	0.03994	< .01	0.00001	< .01	0.01289	0.00003

Table 11. Statistical comparison between the non-preprocessed data set and the resampled data sets. The first line of each method is the *t*-value, and the second line corresponds to the *p*-value (italic values indicate no significant differences, while underline values indicate that the results without resampling were better than those with resampling).

indicated the convenience of over-sampling the normalized data set with the SMOTE algorithm to increase the performance of the prediction models.

As a further confirmation of the findings using SMOTE, in Fig. 8 we plotted precision-recall curves for the best predictiion models (SVM, RF and CatBoost) when applied to the original training sets and the over-sampled training sets. The area under the precision-recall curve was 0.838, 0.860 and 0.872 for SVM, 0.873, 0.91 and 0904 for RF, and 0.872, 0.908 and 0.898 for CatBoost using the original, over-sampled and under-sampled training sets, respectively. These values confirm some performance improvements as a result of addressing class imbalance with SMOTE.

The last experiment focused on analyzing the behavior of the prediction models on the upsized data set using the feature vector with the five most relevant attributes according to the multiple intersection method. Table 12 shows that the best performing models were LR and SVM, which is quite surprising because these results differed from those obtained on the data set containing all attributes. On the other hand, when comparing the results of the upper part of Table 10 with those of Table 12, one can see that the performance of all the prediction models worsened when applied to the reduced sets. To check whether or not the differences were statistically significant, we again ran a two-tailed *t*-test for a significance level of 0.05: t-value $= -6.898545$, p-value $= 0.00098$.

Conclusions

Glioma grading and prediction constitute a highly relevant practical health problem that is usually addressed using neuroimaging techniques. However, the development of advanced genomics and proteomics methods allows the identification of mutations in certain molecular biomarkers that can support diagnosis, prognosis and prediction of response to therapy. In this study, several data-centric machine learning models have been used to discriminate between LGG and GBM samples using a series of clinical factors and molecular biomarkers. Furthermore, a comprehensive descriptive analysis of the data set used in the experiments has also been carried out. The descriptive analysis has included several statistics of the attributes and the application of four feature ranking algorithms to determine the most relevant characteristics, and it has been possible to observe that the molecular biomarkers selected by these algorithms as the most informative agree with the conclusions of previous molecular biology studies. However, these algorithms have important advantages because they are much less expensive and faster than genomics and proteomics methods.

Of the different machine learning methods analyzed, the two classifier ensembles (RF and CatBoost) have obtained the best scores regardless of the metric used. The global feature importance approach revealed the absolute relevance of each attribute, while the SHAP analysis of individual samples provided a reasonable interpretation of which attributes contributed most to the prediction of class GBM. On the other hand, when analyzing the confusion matrices, important differences have been observed between the misclassifications on the majority class and the minority class, which suggested the need to apply some techniques to address the class imbalance. In particular, the normalized data set has been preprocessed with an oversampling algorithm (SMOTE) and an undersampling algorithm (RUS) and it has been found that upsizing the minority class improves the prediction performance. As a final comment, it is worth noting that a model-centric approach applied to the TCGA data

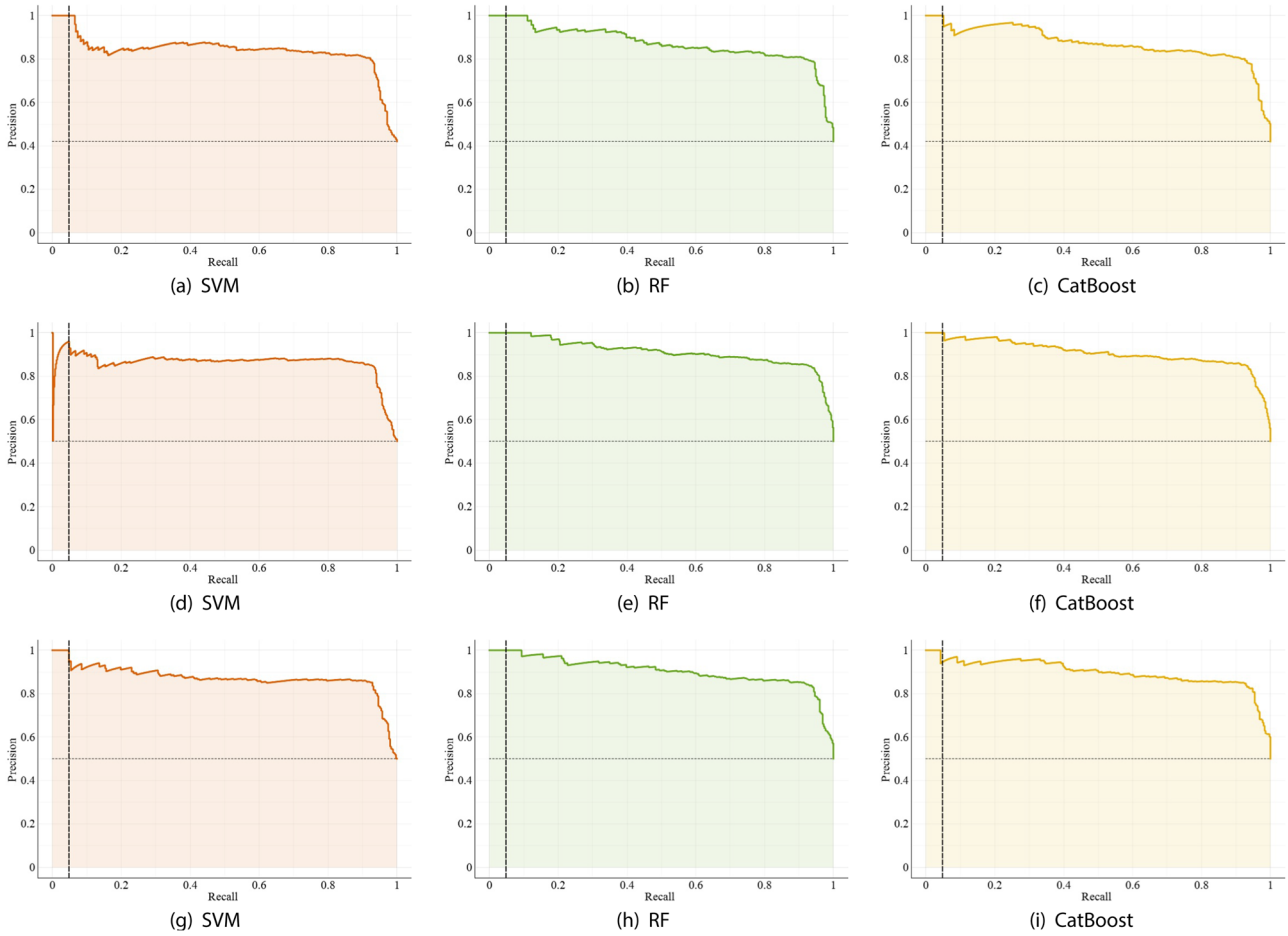

Figure 8. Precision-recall curves for SVM and the classifier ensembles applied with the original training sets (**a**–**c**) , the oversampled training sets (**d**–**f**), and the undersampled training sets (**g**–**i**).

Model	Acc	F1	Prec	Recall	MCC	Spec
kNN	0.853	0.853	0.855	0.853	0.708	0.853
SVM	0.869	0.868	**0.875**	0.869	**0.744**	0.869
MLP	0.868	0.867	0.871	0.868	0.739	0.868
LR	**0.871**	**0.870**	0.873	**0.871**	**0.744**	**0.871**
RF	0.867	0.866	0.871	0.867	0.738	0.867
CatBoost	0.866	0.865	0.870	0.866	0.736	0.866

Table 12. Prediction performance of the machine learning models on the oversampled data set using the top five attributes (the best values are in bold).

set achieved 0.876 accuracy[31], while the data-centric method proposed in this study yielded accuracy rates of 0.882 (with oversampling) and 0.881 (with both oversampling and undersampling).

While this study provides valuable insights into prediction of glioma grades, an interesting avenue for future research refers to the analysis of the possible bias that may arise in predictions against certain sensitive social groups (e.g., gender, age, race, etc.). With this objective, the aim is to quantify the existence of bias through fairness metrics and, if necessary, apply bias mitigation methods[61,62]. When the bias is inherited from the way the training set was created, one approach that would reduce the bias is to internally rebalance the class distributions so that they are equal across class and sensitive attributes.

Data availibility

The data set used and analyzed during the current study is available in the UCI Machine Learning Repository: Glioma Grading Clinical and Mutation Features [Dataset]. https://doi.org/10.24432/C5R62J.

Code availability

The custom code used in this study is available upon request from the corresponding author.

Received: 3 April 2024; Accepted: 22 July 2024
Published online: 26 July 2024

References

1. Louis, D. N. *et al.* The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* **114**, 97–109 (2007).
2. Delgado-López, P. D. & Corrales-García, E. M. Survival in glioblastoma: A review on the impact of treatment modalities. *Clin. Transl. Oncol.* **18**, 1062–1071 (2016).
3. Hanif, F. *et al.* Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical presentation and treatment. *Asian Pac. J. Cancer Prev.* **18**, 3–9 (2017).
4. Zhuge, Y. *et al.* Automated glioma grading on conventional MRI images using deep convolutional neural networks. *Med. Phys.* **47**, 3044–3053 (2020).
5. Kummar, S. & Lu, R. Using radiomics in cancer management. *JCO Precis. Oncol.* **8**, e2400155 (2024).
6. Taha, B., Boley, D., Sun, J. & Chen, C. C. State of radiomics in glioblastoma. *Neurosurgery* **89**, 177–184 (2021).
7. Cheng, J. *et al.* Prediction of glioma grade using intratumoral and peritumoral radiomic features from multiparametric MRI images. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 1084–1095 (2022).
8. Lee, J. H. *et al.* Preoperative prediction of early recurrence in resectable pancreatic cancer integrating clinical, radiologic, and CT radiomics features. *Cancer Imaging* **24**, 6 (2024).
9. Miranda, J. *et al.* The role of radiomics in rectal cancer. *J. Gastrointest. Cancer* **54**, 1158–1180 (2023).
10. Nguyen, H. S. *et al.* Predicting EGFR mutation status in non-small cell lung cancer using artificial intelligence: A systematic review and meta-analysis. *Acad. Radiol.* **31**, 660–683 (2024).
11. Khanfari, H. *et al.* Exploring the efficacy of multi-flavored feature extraction with radiomics and deep features for prostate cancer grading on mpMRI. *BMC Med. Imaging* **23**, 195 (2023).
12. Kim, S., Kim, M. J., Kim, E. K., Yoon, J. H. & Park, V. Y. MRI radiomic features: Association with disease-free survival in patients with triple-negative breast cancer. *Sci. Rep.* **10**, 3750 (2020).
13. Pinter, N. K. & Fritz, J. V. Neuroimaging for the neurologist: Clinical MRI and future trends. *Neurol. Clin.* **38**, 1–35 (2020).
14. Verger, A. & Langen, K. J. PET Imaging in Glioblastoma: Use in Clinical Practice. In *Glioblastoma* (ed. De Vleeschouwer, S.) (Codon Publications, 2017).
15. Almansory, K. O. & Fraioli, F. Combined PET/MRI in brain glioma imaging. *Br. J. Hosp. Med. (Lond.)* **80**, 380–386 (2019).
16. Tiefenbach, J. *et al.* The use of advanced neuroimaging modalities in the evaluation of low-grade glioma in adults: A literature review. *Neurosurg. Focus* **56**, E3 (2024).
17. Siegal, T. Clinical impact of molecular biomarkers in gliomas. *J. Clin. Neurosci.* **22**, 437–444 (2015).
18. Figarella-Branger, D. *et al.* The 2021 WHO classification of tumours of the central nervous system. *Ann. Pathol.* **42**, 367–382 (2022).
19. Zlochower, A. *et al.* Deep learning AI applications in the imaging of glioma. *Top. Magn. Reson. Imaging* **29**, 115–121 (2020).
20. Buchlak, Q. D. *et al.* Machine learning applications to neuroimaging for glioma detection and classification: An artificial intelligence augmented systematic review. *J. Clin. Neurosci.* **89**, 177–198 (2021).
21. Luo, J., Pan, M., Mo, K., Mao, Y. & Zou, D. Emerging role of artificial intelligence in diagnosis, classification and clinical management of glioma. *Semin. Cancer Biol.* **91**, 110–123 (2023).
22. Deepak, S. & Ameer, P. M. Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **111**, 103345 (2019).
23. Shboul, Z. A., Chen, J. & Iftekharuddin, K. M. Prediction of molecular mutations in diffuse low-grade gliomas using MR imaging features. *Sci. Rep.* **10**, 3711 (2020).
24. Alksas, A. *et al.* A novel system for precise grading of glioma. *Bioengineering* **9**, 532 (2022).
25. Matsui, Y. *et al.* Prediction of lower-grade glioma molecular subtypes using deep learning. *J. Neurooncol.* **146**, 321–327 (2020).
26. Gutta, S., Acharya, J., Shiroishi, M. S., Hwang, D. & Nayak, K. S. Improved glioma grading using deep convolutional neural networks. *AJNR Am. J. Neuroradiol.* **42**, 233–239 (2021).
27. Sun, P., Wang, D., Mok, V. C. & Shi, L. Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading. *IEEE Access* **7**, 102010–102020 (2019).
28. Cho, H. H., Lee, S. H., Kim, J. & Park, H. Classification of the glioma grading using radiomics analysis. *PeerJ* **6**, e5982 (2018).
29. Bae, S. *et al.* Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. *Sci. Rep.* **10**, 12110 (2020).
30. Zhao, R., Zhuge, Y., Camphausen, K. & Krauze, A. V. Machine learning based survival prediction in glioma using large-scale registry data. *Health Inform. J.* **28**, 14604582221135428 (2022).
31. Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K. & Krauze, A. V. Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. *Int. J. Mol. Sci.* **23**, 14155 (2022).
32. Joshi, R. C. *et al.* Ensemble based machine learning approach for prediction of glioma and multi-grade classification. *Comput. Biol. Med.* **137**, 104829 (2021).
33. Munquad, S., Si, T., Mallik, S., Li, A. & Das, A. B. Subtyping and grading of lower-grade gliomas using integrated feature selection and support vector machine. *Brief. Funct. Genom.* **21**, 408–421 (2022).
34. Ren, Y. *et al.* Noninvasive prediction of IDH1 mutation and ATRX expression loss in low-grade gliomas using multiparametric MR radiomic features. *J. Magn. Reson. Imaging* **49**, 808–817 (2019).
35. Zheng, S. *et al.* GlioPredictor: A deep learning model for identification of high-risk adult IDH-mutant glioma towards adjuvant treatment planning. *Sc.i Rep.* **14**, 2126 (2024).
36. Zhan, T. *et al.* An automatic glioma grading method based on multi-feature extraction and fusion. *Technol. Health Care* **25**, 377–385 (2017).
37. Wu, M. *et al.* Development and validation of a clinical prediction model for glioma grade using machine learning. *Technol. Health Care* **32**, 1977–1990 (2024).
38. Ye, L. *et al.* An online survival predictor in glioma patients using machine learning based on WHO CNS5 data. *Front. Neurol.* **14**, 1179761 (2023).
39. Zhou, H., Chen, B., Zhang, L. & Li, C. Machine learning-based identification of lower grade glioma stemness subtypes discriminates patient prognosis and drug response. *Comput. Struct. Biotechnol. J.* **21**, 3827–3840 (2023).
40. Kha, Q. H., Le, V. H., Hung, T. N. K. & Le, N. Q. K. Development and validation of an efficient MRI radiomics signature for improving the predictive performance of 1p/19q co-deletion in lower-grade gliomas. *Cancers* **13**, 5398 (2021).
41. Kumar, S., Datta, S., Singh, V., Singh, S. K. & Sharma, R. Opportunities and challenges in data-centric AI. *IEEE Access* **12**, 33173–33189 (2024).
42. Zha, D., Bhat, Z. P. , Lai, K.- H., Yang, F. & Hu, X. Data-centric AI: Perspectives and challenges. In *Proc. SIAM Int. Conf. on Data Mining* (eds Shekhar, S. *et al.*) 945–948 (SIAM, 2023).

43. Hamid, O. H. Data-centric and model-centric AI: Twin drivers of compact and robust industry 4.0 solutions. *Appl. Sci.* **13**, 2753 (2023).
44. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
45. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. In *Proc. 32nd Int. Conf. on Neural Information Processing Systems* (eds Bengio, S. *et al.*) 6639–6649 (ACM, 2018).
46. Yap, B. W. & Sim, C. H. Comparisons of various types of normality tests. *J. Stat. Comput. Simul.* **81**, 2141–2155 (2011).
47. Mishra, P. *et al.* Descriptive statistics and normality tests for statistical data. *Ann. Card. Anaesth.* **22**, 67–72 (2019).
48. DeWitt, J. C. *et al.* Cost-effectiveness of IDH testing in diffuse gliomas according to the 2016 WHO classification of tumors of the central nervous system recommendations. *Neuro-Oncology* **19**, 1640–1650 (2017).
49. Kan, L. K. *et al.* Potential biomarkers and challenges in glioma diagnosis, therapy and prognosis. *BMJ Neurol. Open.* **2**, e000069 (2020).
50. Kruskal, J. B. & Wish, M. *Multidimensional Scaling* (SAGE, 1978).
51. Corani, G. & Benavoli, A. A Bayesian approach for comparing cross-validated algorithms on multiple data sets. *Mach. Learn.* **100**, 285–304 (2015).
52. Gunning, D. *et al.* XAI-Explainable artificial intelligence. *Sci. Robot.* **4**, 120 (2019).
53. Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**, 177 (2019).
54. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) 4765–4774 (Curran Associates, 2017).
55. Alabi, R. O. *et al.* Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Sci. Rep.* **13**, 8984 (2023).
56. López, V., Fernández, A., Moreno-Torres, J. G. & Herrera, F. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Syst. Appl.* **39**, 6585–6608 (2012).
57. García, V., Sánchez, J. S., Marqués, A. I., Florencia, R. & Rivera, G. Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. *Expert Syst. Appl.* **158**, 113026 (2020).
58. Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H. & Santos, J. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier]. *IEEE Comput. Intell. M.* **13**, 59–76 (2018).
59. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
60. Bland, J. M. & Bland, D. G. Statistics notes: One and two sided tests of significance. *BMJ* **309**, 248 (1994).
61. Fletcher, R. R., Nakeshimana, A. & Olubeko, O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front. Artif. Intell.* **3**, 561802 (2021).
62. Giovanola, B. & Tiribelli, S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI Soc.* **38**, 549–563 (2023).

Author contributions
All authors contributed equally to the preparation of this manuscript, read it, and approved the submitted version.

Funding
Open access funding provided by Institute of New Imaging Technologies and Department of Computer Languages and Systems, Universitat Jaume I.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to V.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.