Brazilian Journal of Biometrics (2022), **40**, 371–381
doi:10.28951/bjb.v40i4.634

BRAZILIAN JOURNAL OF BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Classification and biomarker selection in lower-grade glioma using robust sparse logistic regression applied to RNA-seq data

João F. Carrilho[1] and Marta B. Lopes[⋆,1,2,3]

[1]NOVA School of Science and Technology, NOVA University of Lisbon (FCT NOVA), Caparica, Portugal
[2]Center for Mathematics and Applications (NOVA MATH), FCT NOVA
[3]NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), FCT NOVA
[⋆]Corresponding author. Email: marta.lopes@fct.unl.pt

(Received: June 1,2022; Revised: December 14,2022; Accepted: December 14,2022; Published: December 30,2022)

Abstract

Effective diagnosis and treatment in cancer is a barrier for the development of personalized medicine, mostly due to tumor heterogeneity. In the particular case of gliomas, highly heterogeneous brain tumors at the histological, cellular and molecular levels, and exhibiting poor prognosis, the mechanisms behind tumor heterogeneity and progression remain poorly understood.

The recent advances in biomedical high-throughput technologies have allowed the generation of large amounts of molecular information from the patients that combined with statistical and machine learning techniques can be used for the definition of glioma subtypes and targeted therapies, an invaluable contribution to disease understanding and effective management.

In this work sparse and robust sparse logistic regression models with the elastic net penalty were applied to glioma RNA-seq data from The Cancer Genome Atlas (TCGA), to identify relevant transcriptomic features in the separation between lower-grade glioma (LGG) subtypes and identify putative outlying observations. In general, all classification models yielded good accuracies, selecting different sets of genes. Among the genes selected by the models, *TXNDC12*, *TOMM20*, *PKIA*, *CARD8* and *TAF12* have been reported as genes with relevant role in glioma development and progression. This highlights the suitability of the present approach to disclose relevant genes and fosters the biological validation of non-reported genes.

Keywords: Glioma; Classification; Sparse Logistic regression; Robust Statistics; Elastic net regularization.

1. Introduction

Lower-grade gliomas (LGG) represent a group of tumors of the Central Nervous System (CNS) arising from the supporting glial cells of the CNS (Youssef & Miller, 2020). The highly invasive nature and incomplete surgical resection in LGG are major responsible for tumor recurrence and progression into high-grade gliomas, namely glioblastoma (GBM) (Kang *et al.,* 2021). LGG are heterogeneous tumors at the histopathological and genotypic levels (Louis *et al.,* 2021), which calls for the need to disclose diagnostic biomarkers and therapeutic targets towards the improvement in the procedures applied to each patient, mainly diagnosis and treatment.

The great advances in the development of technologies such as high-throughput screening and mass spectrometry, globally designated as "omics", have now made possible getting into the molecular heterogeneity of tumors and better characterizing cancer subtypes. The classification of LGG subtypes has been evolving as the understanding of tumors progresses. The latest classification (Louis *et al.,* 2021; WHO, 2021) introduces changes that reflect the increasing role of molecular features in complementing other established approaches to tumor characterization, namely histology and immunohistochemistry, some based on novel technologies such as DNA methylomics (Louis *et al.,* 2021).

Despite the massive amounts of data generated by omics' technologies, identifying the most relevant information out of these high-dimensional data remains a complex task. The application of machine learning techniques has shown promising in multi-omics data analysis, handling well with the complexity of biological data in order to produce efficient results (Cai *et al.,* 2022). In the context of gliomas, machine learning techniques are useful for extracting relevant biomarkers to support treatment decisions and patient monitoring (Wu *et al.,* 2021).

One of the major challenges associated with the high-dimensional nature of omics data is the need to find lower dimensions of the data which are informative, and to identify the most relevant features in the molecular structure underlying the disease development and progression. Dimensionality reduction and feature selection combined with pattern recognition methods have been used for that purpose. Among these techniques, logistic regression and its sparsity-inducing modifications have been widely used in biomedical research, and their ability to dealing with several data types recently led to its even more frequent application (Hastie *et al.,* 2015). In gliomas' research, sparse logistic regression using different regularizers was successfully applied for classifying glioma subtypes and grades, and GBM cell clones based on gene expression and radiomic features (Liu *et al.,* 2008; Lopes & Vinga, 2020; Nakamoto *et al.,* 2019).

A wrong diagnosis leads to severe consequences regarding the appropriateness of the therapy prescribed, and the overall cancer progression and survival. Therefore, models should be accurate in classifying patients in the correct disease class and at the same time being able to identify patients deviating from the overall pattern of their attributed class. These might reflect a wrong class membership or point to outlying features that deserve further investigation at the individual and molecular levels. In high-dimensional scenarios, as in the case of omics data, with far more variables than observations, both classical methods and sparsity-inducing counterparts are highly influenced by these outlying observations, not being able to detect them and ultimately rendering regular observations as outliers, also known as the masking and swamping effects (Serfling & Wang, 2014). Outlier observations not only strongly impact parameter estimation but also variable selection (Alfons *et al.,* 2013). Therefore, methods that are robust to observations that deviate from the remaining observations in the same group have been proposed, in particular, sparse modifications of the popular Least Trimmed Squares (LTS) robust estimator (Rousseeuw, 2013; Rousseeuw & Driessen, 2006) have been successful applied to high-dimensional data, namely gene expression cancer data (Alfons *et al.,* 2013; Jensch *et al.,* 2022; Segaert *et al.,* 2019; Sun *et al.,* 2021).

The goal of this work is the identification of key transcriptomic markers in LGG through sparse logistic regression. Robust methods are used to classify patients into astrocytoma and oligodendroglioma LGG subtypes, identify the relevant gene features separating the patient groups, and identify outlying patients whose transcriptomic profile differs from the members of the same subtype. The LGG RNA-sequencing (RNA-seq) dataset was obtained from The Cancer Genome Atlas (TCGA) data portal[1].The obtained results are expected to contribute to glioma disease understanding, therapy research and disease management.

2. Materials and Methods

2.1 Dataset

The glioma RNA-seq dataset was extracted from The Cancer Genome Atlas (TCGA) data portal, comprising the expression of 20501 variables from 659 patients with the following glioma subtypes: glioblastoma (GBM, 149 patients), astrocytoma (LGG-a, 193 patients), oligoastrocytoma (LGG-oa, 129 patients) and oligodendroglioma (LGG-od, 188 patients). Further classification analysis was performed based on LGG patients, belonging to LGG-a and LGG-od subtypes, excluding LGG-oa, a LGG subtype showing histological and molecular characteristics of both LGG-a and LGG-od subtypes, therefore considered a mixed subtype (Sahm *et al.,* 2014).

2.2 Binary classification methods

2.2.1 Sparse logistic regression with the elastic net penalty (SLR)

Let $X = \{x_1, \dots, x_N\}$ be the set of observations, with each x_i having $p \in \mathbb{N}$ variables and an outcome $y_i \in Y$ that can be 0 or 1. In this work, class 1 represents patients with LGG-a, with class 0 representing patients with LGG-od. Logistic regression considers that the logit function of the probabilities of the observations x_i belonging to class 1 can be modeled with a linear regression as follows:

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \boldsymbol{\beta}^\top x_i + \varepsilon_i, \quad \boldsymbol{\beta} = (\beta_1, \dots, \beta_p), \quad \pi_i = P(Y_i = 1|X_i = x_i), \tag{1}$$

with β_j corresponding to the coefficient associated to j-th variable, $j \in \{1, \dots, p\}$, and β_0 the independent term of the regression. Equation (1) leads to the modeling of those probabilities with a sigmoid function with $p + 1$ parameters, defined as

$$P(Y_i = 1|X_i = x_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^\top x_i)}}. \tag{2}$$

The predicted probability indicates which class an observation will be assigned to, choosing class 1 for x_i if $\pi_i \geq 0.5$ and class 0 otherwise. The β_0 and $\boldsymbol{\beta}$ coefficients in Equation (2) are then estimated by minimizing the penalized negative log-likelihood function (Hastie *et al.,* 2015) as

$$l(\beta_0, \boldsymbol{\beta}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \right] + \lambda F_\alpha(\boldsymbol{\beta}), \tag{3}$$

where $F_\alpha(\boldsymbol{\beta})$ induces sparsity in the classifier and represents the regularization term, which for the elastic net (Zou & Hastie, 2005) is defined as

$$F_\alpha(\boldsymbol{\beta}) = \sum_{i=1}^{p} \left(\alpha|\beta_i| + \frac{1-\alpha}{2} \beta_i^2 \right), \tag{4}$$

with α controlling the balance between the Lasso and Ridge penalties and λ controlling the strength of the penalty.

[1] https://www.cancer.gov/tcga.

2.2.2 Robust sparse logistic regression with the elastic net penalty (rSLR)

Although SLR solves the problem of high dimensionality of the data inducing sparsity, these class of models is still limited regarding outliers, since this observations aren't penalized, biasing the optimum values obtained with Equation (3).

To handle this common issue, an adaptation of the Least Trimmed Squares (LTS) estimator for robust linear regression was introduced in Kurnaz *et al.* (2018), given by

$$Q(H, \beta_0, \boldsymbol{\beta}) = -\frac{1}{h}\sum_{i\in H}\left[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\right] + \lambda F_\alpha(\boldsymbol{\beta}), \tag{5}$$

where $H \subseteq \{1, 2, \ldots, N\}$ with $|H| = h$ and $F_\alpha(\boldsymbol{\beta})$ is defined in Equation (4).

Therefore, β_0 and $\boldsymbol{\beta}$, as well as an optimal subset H of indexes, are now estimated by minimizing $Q(H, \beta_0, \boldsymbol{\beta})$. In this work, h corresponds to 75% of the number of observations in the training set (see Section 2.3).

Given its low efficiency, a reweighting step is commonly used to improve the LTS estimator (Rousseeuw & Leroy, 2005). In this step, the Pearson residuals, given by

$$r_i^s = \frac{y_i - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}, \quad i \in \{1, \ldots, N\}, \tag{6}$$

with π_i defined in Equation (1), are computed since they are approximately normally distributed for the logistic model.

The outliers of the current classifier are identified and reweighted through the application of a weight w_i to each observation x_i, defining $w_i = 1$ if $\left|r_i^s\right| \leq \Phi^{-1}(0.975)$ and $w_i = 0$ otherwise, with Φ^{-1} being the inverse of the cumulative distribution function of the standard Gaussian distribution. This way, the model flags 2.5% of the observations as outliers.

The function to minimize in order to obtain the values for β_0 and $\boldsymbol{\beta}$ in the reweighted model becomes

$$Q_r(\beta_0, \boldsymbol{\beta}) = -\frac{1}{N_w}\sum_{i=1}^{n} w_i\left[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\right] + \lambda_{\text{upd}} F_\alpha(\boldsymbol{\beta}), \tag{7}$$

with $N_w = \sum_{i=1}^{N} w_i$ and λ_{upd} the update of the λ value for the reweighted model, obtained by cross-validation with the value of α already fixed.

2.3 Model construction

Before the application of the classification methods in Section 2.2, the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes *et al.*, 2018) was used to visualize LGG observations in a low-dimensional space and to get an initial idea of the separability between both classes. UMAP is a popular non-linear feature extraction algorithm which allows the visualization of groups of samples from high-dimensional data, while preserving of the global structure of the data.

For the construction of the classification models, the dataset was split into training (75%) and test (25%) sets, leaving 285 patients for training and 96 patients for testing, and each observation standardized by subtracting by the training set's sample mean value and dividing by the training set's standard deviation for each feature.

Several SLR models were generated, using 8-fold cross-validation for the tuning of the parameter λ by optimization of $l(\beta_0, \boldsymbol{\beta})$ (Equation (3)), considering $\alpha \in \{0, 0.1, 0.2, ..., 0.9, 1\}$.

Following the same data splitting approach, cross-validation was performed to find the pair (α, λ) that minimizes, in average, $Q(H, \beta, \boldsymbol{\beta})$ (Equation (5)). The resulting parameter set was used to build a rSLR model. All models were evaluated regarding the performance measured by the area under the receiver operating characteristic (ROC) curve (AUC), the number of misclassifications, and the number of features selected.

The methods described in sections 2.1 and 2.2 were implemented and tested using version 4.1.3 of the R software (R Core Team, 2022), using packages `umap` (Konopka, 2022), `readr` (Wickham *et al.*, 2022), `glmnet` (Friedman *et al.*, 2010), `enetLTS` (Kurnaz *et al.*, 2022), `pROC` (Robin *et al.*, 2011) and `VennDiagram` (Chen, 2022).

3. Results and Discussion

3.1 Classification

A first exploratory analysis was performed to visualize how known histological LGG groups are distributed in a lower dimensional UMAP space generated from gene expression data.

In Figure 1 some grouping structure can be found, with clusters of LGG-a (green) and LGG-od (purple) observed, despite showing some overlap. This result highlights the relevance of transcriptomic data to separate LGG subtypes and supports its further use for LGG classification and biomarker selection, as discussed next.

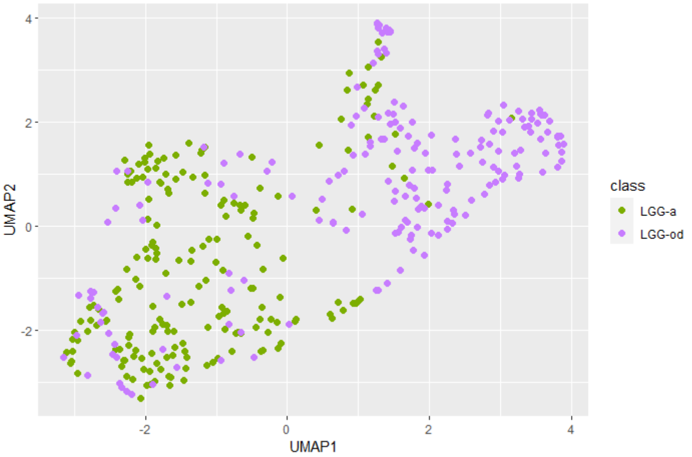

Figure 1. UMAP representation of the dataset (LGG-a: patients with astrocytoma; LGG-od: patients with oligodendroglioma).

Prediction of LGG subtypes was then performed by SLR based on the RNA-seq data. Overall, good prediction models were obtained in both training and test sets for all α values evaluated, regarding the number of misclassifications and AUC values (Table 1). The best predictive performance was obtained for $\alpha = 0$, yielding an AUC value of 0.8107 in the test set, though selecting almost the entire gene set. As the goal is the selection of a small set of interpretable features, other α values might be considered, e.g., $\alpha = 0.5$, which selects a lower number of features and performs similarly in the test set.

LGG subtypes were then predicted by rSLR. The λ values in Figure 2 were chosen according to the λ_{1se} range that was obtained for SLR. The (α, λ) pair minimizing the loss function for rSLR was obtained for $\alpha_{\text{opt}} = 0.7$ and $\lambda_{\text{opt}} = 0.02$. The parameters were chosen for building the rSLR model, whose results are summarized in Table 2.

Comparable predictive performance was obtained for rSLR with respect to SLR for the range of α values considered. A total of 23 observations were flagged as outliers by rSLR (3 for the LGG-a class and 20 for the LGG-od class), ending up as misclassifications in the training set. For some of

Table 1. Results obtained with each fitted SLR (λ_{1se}: the largest λ at which the mean squared error (MSE) is within one standard error of the smallest MSE in the cross-validation; $\beta_i \neq 0$: number of features selected; Misc.: number of misclassifications; AUC: area under the ROC curve value).

α	λ_{1se}	$\beta_i \neq 0$	Training set		Test set	
			Misc.	AUC	Misc.	AUC
0	88.3081	20140	34	0.8799	18	0.8107
0.1	3.1007	7	130	0.5390	42	0.5532
0.2	0.7031	79	41	0.8552	19	0.8000
0.3	0.4910	52	40	0.8587	19	0.8000
0.4	0.3203	42	39	0.8623	19	0.8000
0.5	0.2127	41	36	0.8729	19	0.8005
0.6	0.1692	31	36	0.8729	20	0.7898
0.7	0.1385	30	36	0.8729	20	0.7898
0.8	0.1054	30	36	0.8729	22	0.7694
0.9	0.0894	26	36	0.8729	23	0.7592
1	0.0805	22	37	0.8693	23	0.7592

these outlying observations, their proximity to the opposite class can be confirmed in the UMAP dimensions generated considering the expression data of the 99 genes selected by the rSLR model (Figure 3), where the separation between the two LGG classes becomes more evident.

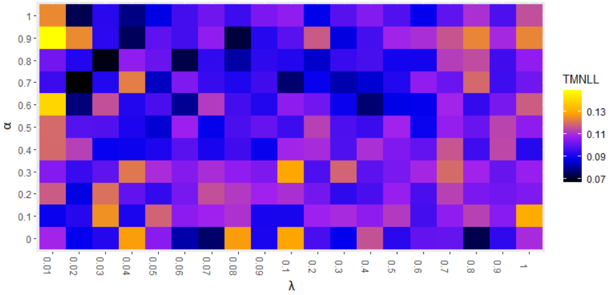

Figure 2. Cross-validation in order to find the pair (α, λ) that minimizes the average loss function for rSLR (TMNLL: mean of the negative log-likelihoods).

Table 2. Results obtained with the fitted rSLR (α_{opt} and λ_{opt}: the pair (α, λ) that gives the optimum value for the loss function in the cross validation; λ_{upd}: the λ value for the reweighted model; $\beta_i \neq 0$: number of features selected; Misc.: number of misclassifications; AUC: area under the ROC curve value).

α_{opt}	λ_{opt}	λ_{upd}	$\beta_i \neq 0$	Training set		Test set	
				Misc.	AUC	Misc.	AUC
0.7	0.02	0.0171	99	27	0.9046	21	0.7792

3.2 Selected genes

For biological interpretation purposes, a closer inspection on the genes selected by SLR and rSLR was performed. Considering the genes selected in common by the SLR models excluding the one obtained with $\alpha = 0.1$ due to its low AUC value for the test set (see Table 3), the thioredoxin domain-containing 12 (*TXNDC12*) gene, which is a highly expressed gene in gliomas and has been pointed as a potential molecular marker for glioma pathological grade and prognosis (Wang *et al.*, 2021), showed one of the largest absolute coefficient values.

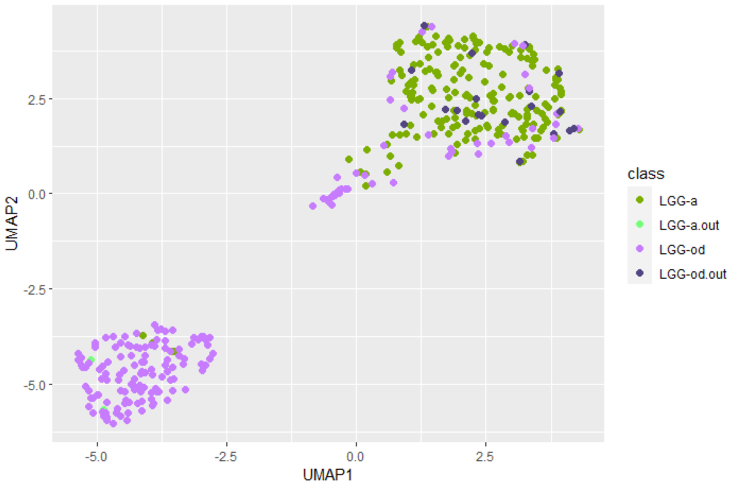

Figure 3. UMAP representation of the dataset based on the features selected by the rSLR model (LGG-a: patients with astrocytoma; LGG-a.out: outlier patients within the astrocytoma class; LGG-od: patients with oligodendroglioma; LGG-od.out: outlier patients within the oligodendroglioma class).

Table 3. Genes selected in common by the SLR models (excluding $\alpha = 0.1$), ordered left to right by decreasing absolute values of the coefficients in the SLR model obtained for $\alpha = 0.2$.

LSM14A	*CARD8*	*TXNDC12*	*NADK*	*THRAP3*
TRAPPC3	*FAM155A*	*LRRTM4*	*WLS*	*VRK3*
USF2	*LOC148189*	*ERCC1*	*ACADM*	*XRCC1*
PITPNB	*PRAM1*	*FGF20*		

The intersection of the gene sets obtained by SLR and rSLR models is illustrated in Figure 4. The SLR model obtained for $\alpha = 0.2$ was chosen for its similarity in the number of selected features and comparable AUC value with the rSLR model. Thirty-three genes were selected in common by the two models, leaving 46 exclusively selected by SLR and 66 by rSLR.

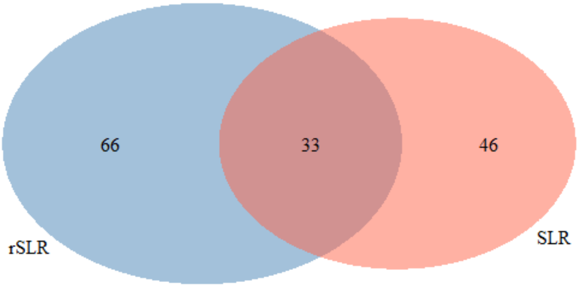

Figure 4. Venn diagram with the intersection of the genes selected by SLR with $\alpha = 0.2$ and rSLR.

The genes selected by the models can be found in Table 4. Among the relevant genes identified in the separation of patients into LGG-a and LGG-od subtypes, several have been reported in the literature as playing a role in gliomas, with the translocase of outer mitochondrial membrane 20 (*TOMM20*) (Zheng *et al.*, 2022), cAMP-dependent protein kinase inhibitor alpha (*PKIA*) (Ratushna, 2020), caspase recruitment domain-containing protein 8 (*CARD8*) (Sharma *et al.*, 2019) and TATA-box binding protein associated factor 12 (*TAF12*) (Ren *et al.*, 2015; Wijethilake *et al.*, 2020) standing out as genes involved in glioma proliferation and tumorigenesis. These genes also showed among the largest coefficients for rSLR (Figure 5).

Table 4. Genes selected exclusively and simultaneously by SLR and rSLR (highlighted genes have their rSLR's absolute coefficients greater than 0.15; *Ordered left to right by decreasing absolute values of the coefficients in rSLR; **Ordered left to right by decreasing absolute values of the coefficients in SLR).

genes(rSLR) \ genes(SLR)*	***TOMM20***	***FNBP1L***	***MLLT3***	***PKIA***	***SPINK5***
	GRIK5	*IQCF6*	*AKT2*	*PTPN20A*	*FAM13C*
	CCL22	*GRM8*	*MTF2*	*CASC3*	*TCHH*
	LEKR1	*ZNF792*	*WNT9B*	*ZMYM4*	*TTLL12*
	GOLGA6L1	*CNTNAP2*	*LOC401463*	*SEC63*	*TTC19*
	OR11H12	*ADH6*	*ZNF766*	*SHANK2*	*GPR119*
	C1orf50	*C15orf57*	*DGCR2*	*ZNF813*	*TARDBP*
	FARP1	*TMEM30A*	*RPAP2*	*SLC25A44*	*ZFAND6*
	FMOD	*CCDC23*	*FAM134A*	*LRPPRC*	*ARV1*
	OSBPL9	*OR5K2*	*OR13H1*	*C9orf93*	*THEM4*
	PKN2	*IL1F6*	*MMP26*	*DEFB110*	*SYN3*
	PEF1	*MYL10*	*LSG1*	*AFARP1*	*ZNF71*
	MCF2	*STK38*	*DBT*	*SNX7*	*KAZ*
	NLRP3				
genes(SLR) \ genes(rSLR)**	*TXNDC12*	*NADK*	*THRAP3*	*AK2*	*FAM155A*
	WLS	*ERCC1*	*CHGB*	*HDAC1*	*XRCC1*
	WDR77	*RIC3*	*CSDE1*	*MOV10*	*SF3A3*
	ETHE1	*PITPNB*	*LOC148413*	*ZNF362*	*PRAM1*
	CDK11B	*INPP5D*	*CAPZB*	*NECAP2*	*SCP2*
	SLAIN1	*DNAJC8*	*ATCAY*	*PRKD2*	*STAT5A*
	RHOC	*PSMC4*	*RER1*	*CACNG2*	*FAM54B*
	SNRNP40	*LSM10*	*PHACTR4*	*ASAP3*	*U2AF2*
	PAFAH2	*NFIA*	*MEGF8*	*ZDHHC22*	*PEPD*
	SDF4				
genes(SLR) ∩ genes(rSLR)*	***LOC148189***	***CARD8***	***CD3EAP***	***ZNF691***	***TTC4***
	VRK3	***TAF12***	*LSM14A*	*FGF20*	*PABPC4*
	USF2	*WASF2*	*FBXO42*	*POP4*	*C12orf43*
	TMEM87A	*S100PBP*	*LRRTM4*	*ACADM*	*ZNF181*
	MIER1	*SFRS4*	*KDELR1*	*GPBP1L1*	*C19orf61*
	PAK4	*TRAPPC3*	*ADPRHL2*	*KIAA2013*	*NUP62*
	HECTD3	*PDCD2L*	*CMPK1*		

4. Conclusions

Understanding heterogeneity in gliomas is a critical step towards the definition of appropriate diagnostic and therapy decision that ultimately will help improving patient prognosis. Biomedical research now benefits from the advances in high-throughput technologies generating high-dimensional omics data. Despite the remarkable advances in the molecular understanding of tumors, uncertainty remains for tumors deviating from the overall pattern of the defined tumor subtypes. This fosters the need to develop appropriate statistical and machine learning strategies which are able to identify the relevant features that accurately predict tumor subtype and identify outlying observations. Such approach might be particularly relevant in the case of gliomas, for which great

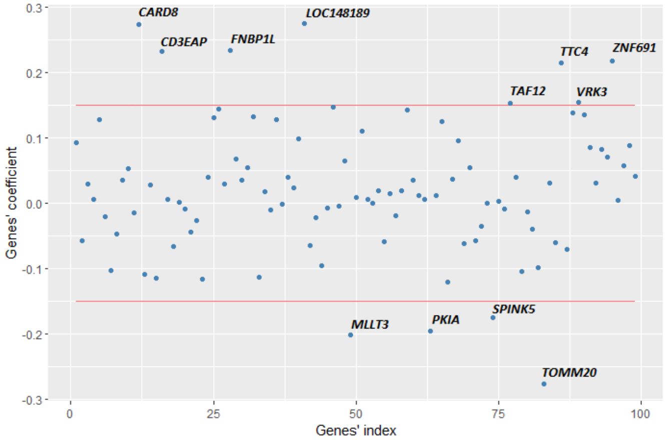

Figure 5. Values for the coefficients of the genes that were selected by the rSLR classifier.

efforts have been pursued to incorporate molecular information into classification guidelines. In this work, sparse and robust sparse logistic regression was applied to LGG RNA-seq data to predict LGG subtypes and identify outlying patients. Despite the overlap of the different subtypes, both SLR and rSLR performed well in their separation. Several observations were flagged as outliers by rSLR and at the same time classified in the opposite class. Further efforts might be pursued to confirm patient outlierness through the update of patient labels according to most recent CNS classification by WHO. Moreover, among the genes selected as relevant in the separation between LGG subtypes, several are known to have a role in cancer and in gliomas (e.g, *TXNDC12*, *TOMM20*, *PKIA*, *CARD8* and *TAF12*), therefore encouraging further biological validation of genes with an unknown role in gliomas among the genes selected.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references CEECINST/00102/2018, UIDB/00297/2020 and UIDB/00297/2020 (NOVA MATH, Center for Mathematics and Applications), UIDB/04516/2020 (NOVA LINCS), and the research project "MONET – Multi-omic networks in gliomas" (PTDC/CCI-BIO/4180/2020). The results presented are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Alfons, A., Croux, C. & Gelper, S. Sparse Least Trimmed Squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics* **7,** 226–249 (2013).

2. Cai, Z., Poulos, R. C., Liu, J. & Zhong, Q. Machine learning for multi-omics data integration in cancer. *iScience* **25,** 103798 (2022).

3. Chen, H. *VennDiagram: Generate High-Resolution Venn and Euler Plots* R package version 1.7.3 (2022). https://CRAN.R-project.org/package=VennDiagram.

4. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33,** 1–22. https://www.jstatsoft.org/v33/i01/ (2010).

5. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations* 143 (CRC Press, 2015).

6. Jensch, A., Lopes, M. B., Vinga, S. & Radde, N. ROSIE: Robust sparse ensemble for outlier detection and gene selection in cancer omics data. *Statistical Methods in Medical Research* **31,** 947–958 (2022).

7. Kang, K., Xie, F., Wu, Y., Han, C., Bai, Y., Long, J., Lian, X. & Zhang, F. Genomic instability in lower-grade glioma: Prediction of prognosis based on lncRNA and immune infiltration. *Molecular Therapy - Oncolytics* **22,** 431–443 (2021).

8. Konopka, T. *umap: Uniform Manifold Approximation and Projection* R package version 0.2.8.0 (2022). https://CRAN.R-project.org/package=umap.

9. Kurnaz, F. S., Hoffmann, I. & Filzmoser, P. *enetLTS: Robust and Sparse Methods for High Dimensional Linear and Binary and Multinomial Regression* R package version 1.1.0 (2022). https://CRAN.R-project.org/package=enetLTS.

10. Kurnaz, F. S., Hoffmann, I. & Filzmoser, P. Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems* **172,** 211–222 (2018).

11. Liu, Z., Gartenhaus, R. B., Tan, M., Jiang, F. & Jiao, X. Gene and pathway identification with L_p penalized Bayesian logistic regression. *BMC Bioinformatics* **412** (2008).

12. Lopes, M. B. & Vinga, S. Tracking intratumoral heterogeneity in glioblastoma via regularized classification of single-cell RNA-Seq data. *BMC Bioinformatics* **21,** 59 (2020).

13. Louis, D. N. *et al.* The 2021 WHO classification of tumors of the Central Nervous System: a summary. *Neuro-Oncology* **23,** 1231–1251 (2021).

14. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

15. Nakamoto, T. *et al.* Prediction of malignant glioma grades using contrast-enhanced T1-weighted and T2-weighted magnetic resonance images based on a radiomic analysis. *Scientific Reports* **9** (2019).

16. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2022). https://www.R-project.org/.

17. Ratushna, O. O. Glucose deprivation affects the expression of genes encoding cAMP-activated protein kinase and related proteins in U87 glioma cells in ERN1 dependent manner. *Endocrine Regulations* **54,** 244–254 (2020).

18. Ren, J., Lou, M., Shi, J., Xue, Y. & Cui, D. Identifying the genes regulated by IDH1 via gene-chip in glioma cell U87. *International journal of clinical and experimental medicine* **8,** 18090 (2015).

19. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (2011).

20. Rousseeuw, P. J. Least median of squares regression. *Journal of the American Statistical Society* **79,** 971–880 (2013).

21. Rousseeuw, P. J. & Driessen, K. V. Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery* **12,** 29–45 (2006).

22. Rousseeuw, P. J. & Leroy, A. M. *Robust regression and outlier detection* (John wiley & sons, 2005).

23. Sahm, F. *et al.* Farewell to oligoastrocytoma: in situ molecular genetics favor classification as either oligodendroglioma or astrocytoma. *Acta Neuropathologica* **128,** 551–559 (2014).

24. Segaert, P, Lopes, M. B., Casimiro, S, Vinga, S & Rousseeuw, P. J. Robust identification of target genes and outliers in triple-negative breast cancer data. *Statistical Methods in Medical Research* **28,** 3042–3056 (2019).

25. Serfling, R. & Wang, S. General foundations for studying masking and swamping robustness of outlier identifiers. *Statistical Methodology* **20,** 79–90 (2014).

26. Sharma, N., Saxena, S., Agrawal, I., Singh, S., Srinivasan, V., Arvind, S., Epari, S., Paul, S. & Jha, S. Differential Expression Profile of NLRs and AIM2 in Glioma and Implications for NLRP12 in Glioblastoma. *Scientific Reports* **9,** 8480 (2019).

27. Sun, H, Wang, J, Zhang, Z, Hu, N & Wang, T. An Efficient Algorithm for the Detection of Outliers in Mislabeled Omics Data. *Computational and Mathematical Methods in Medicine* **9436582** (2021).

28. Wang, X., Yang, Q., Liu, N., Bian, Q., Gao, M. & Hou, X. Clinical Value of TXNDC12 Combined With IDH And 1p19q as Biomarkers for Prognosis of Glioma. *Pathology & Oncology Research* **27,** 1609825 (2021).

29. WHO. WHO Classification of Tumours Editorial Board. World Health Organization Classification of Tumours of the Central Nervous System. *5th ed. Lyon: International Agency for Research on Cancer* (2021).

30. Wickham, H., Hester, J. & Bryan, J. *readr: Read Rectangular Text Data* R package version 2.1.2 (2022). https://CRAN.R-project.org/package=readr.

31. Wijethilake, N., Meedeniya, D., Chitraranjan, C. & Perera, I. *Survival prediction and risk estimation of Glioma patients using mRNA expressions* in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)* (2020), 35–42.

32. Wu, Y., Guo, Y., Ma, J., Sa, Y., Li, Q. & Zhang, N. Research Progress of Gliomas in Machine Learning. *Cells* **10,** 3169 (2021).

33. Youssef, G. & Miller, J. J. Lower Grade Gliomas. *Current Neurology and Neuroscience Reports* **20** (2020).

34. Zheng, J., Zhou, Z., Qiu, Y., Wang, M., Yu, H., Wu, Z., Wang, X. & Jiang, X. A Pyroptosis-Related Gene Prognostic Index Correlated with Survival and Immune Microenvironment in Glioma. *Journal of Inflammation Research* **15,** 17–32 (2022).

35. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* **67,** 301–320 (2005).