



XII Iberian Modeling Week

Predicting brain tumor subtypes from gene expression data

Coordinator: Dr. Marta Lopes, NOVA School of Science and Technology, Portugal

Work carried out by:

Catarina Marques

Enrique Rodríguez

Isabel Álvares Ribeiro

Vitor Saramago

Abstract

In the context of the Iberian Modeling Week event, this report was developed to highlight the application of statistics and machine learning models to a critical challenge: distinguishing between two subtypes of low-grade glioma (oligodendroglioma and astrocytoma) using gene expression data (RNA-seq). Given the high dimensionality of the data, we implemented techniques namely Random Forest and Logistic Regression with LASSO regularization, using the R programming language for data modeling and visualization. The results demonstrate that both models can effectively distinguish between the tumor subtypes. Regarding the identification of biomarkers, the LASSO model successfully accomplished the feature selection objective, isolating a reduced set of genes. On the other hand, as variable selection is not the primary goal of the Random Forest algorithm, we relied on feature importance analysis, which allowed us to extract the genes with the greatest weight in the model. The comparison between the two techniques revealed distinct groups of genes, reflecting the different statistical natures of the algorithms when dealing with biological variables exhibiting similar behavior.

Contents

Introduction	2
1 Data preprocessing	4
1.1 Variance Filtering	4
1.2 Exploratory Analysis	5
1.3 Data split	6
2 Random Forest Method	7
2.1 Fundamentals of Decision Trees	7
2.2 Model Configuration and Splitting Criteria	8
2.3 Error Metrics and Purity Measures	8
2.4 Confusion Matrices	9
2.5 Strengths and Limitations	9
3 Logistic Regression Model	10
4 Sparse Logistic Regression with L1 Regularization	11
4.1 The Elastic Net and LASSO Approach	11
4.2 Cross-Validation and Parameter Selection	12
4.3 Confusion Matrices	12
4.4 Strengths and Limitations	13
5 Model Comparison and Gene Signature Analysis	14
5.1 Performance LASSO vs Random Forest	14
5.2 Post-selection Visualization	16
6 Conclusion	18
7 References	19

Introduction

In this work, our objective was to translate the biological complexity of cancer into mathematical and computational language, analyzing how machine learning algorithms can assist in medical diagnosis. Taking this into account, bioinformatics and data science emerge as highly relevant fields, providing tools capable of handling the high dimensionality of genomic data.

Clinical Context

This work is part of the Iberian Modelling Week event, and its main objective is to illustrate the application of machine learning models in distinguishing brain tumor subtypes. The focus of this study is on two subtypes of low-grade gliomas: Oligodendroglioma and Astrocytoma.

Objectives

- To build models that can accurately classify patients into the correct tumor subtype.
- To identify a small set of genes (biomarkers) that are determinant for discriminating between the tumor subtypes.
- To apply data preprocessing techniques to mitigate noise and genomic information redundancy.

Dataset Description and Challenges

Correctly distinguishing between these two subtypes is critical for improving patient prognosis and clinical treatment. In this project, to uncover genetic patterns that enable this distinction, we utilized RNA sequencing (RNA-seq) data from The Cancer Genome Atlas (TCGA). The major challenge of this project lies in the nature of the data: the number of variables (20 501) is larger than the number of observations (420 patients), a statistical phenomenon known as the curse of dimensionality ($p \gg n$). Adding to this phenomenon is the problem of multicollinearity, as several genes exhibit highly correlated expressions. Furthermore, the presence of thousands of genes irrelevant to tumor distinction acts as statistical noise, which can lead to overfitting. If this event is not avoided, it causes the model to memorize the statistical noise instead of learning the true genetic patterns, severely affecting prediction efficacy in new patients.

As one of the objectives is to classify patients according to their low-grade glioma subtype, we assigned the following binary classification:

$$Y(X) = \begin{cases} 0, & \text{if Astrocytoma,} \\ 1, & \text{if Oligodendroglioma.} \end{cases}$$

where $X \in \mathbb{R}^{n \times p}$ is the matrix of predictors, with $n = 420$ being the number of patients and $p = 20501$ the number of genes.

Project Development and Structuring

The work developed is based on the premise that gene expression contains patterns capable of differentiating the two glioma subtypes under study. Initially, a filtering step was performed: genes with variance below the 10th percentile were eliminated, leaving the remaining 18 451 variables. Next, we conducted an exploratory data analysis using Principal Component Analysis (PCA) and Uniform Manifold Approximation (UMAP) techniques. Subsequently, to ensure effective model learning rather than answer memorization, the data was split into training and test sets. The training set contains 75% of the data ($n = 305$), and the testing set contains the remaining 25% of the data ($n = 105$).

During the modeling phase, we first tested the Random Forest algorithm. This model proved to be highly effective when evaluated by the confusion matrix, accuracy, and AUC (Area Under the ROC Curve), and we were able to extract the 26 most important genes through feature importance. However, given this model's limitation in executing direct gene selection, we moved on to the logistic model. The traditional logistic model presented several flaws when dealing with high-dimensional data. In order to overcome them, this model was refined with the LASSO penalty. The logistic regression model with LASSO regularization demonstrated its efficacy through the same evaluation methods used for the Random Forest (confusion matrix, accuracy and AUC) and successfully resolved the previous problem related to gene selection.

Motivation and Acquired Skills

The main motivation for this project was to understand how statistics and machine learning can be used to solve real-world problems in healthcare. The work provided the development of technical and scientific skills, notably the manipulation of high-dimensional data, the implementation of classification algorithms in the R programming language, and the critical analysis of computational results against biological reality.

Report Organization

The report contains an abstract and an introduction, and is divided into 6 main chapters: **Chapter 1** details data preprocessing, namely dimensionality reduction through variance filtering, presents the exploratory analysis using techniques such as PCA and UMAP, and describes the split of the data into a training set and a test set. Following this, **Chapter 2** focuses on the implementation of the Random Forest method. **Chapter 3** addresses the logistic regression model, explaining its limitations and the reason why it was not possible to progress with this model. To overcome this, **Chapter 4** presents the sparse logistic regression with L1 regularization, detailing the elastic net and LASSO approach as a dual solution for glioma classification and gene selection. **Chapter 5** is focused on comparing the efficacy and performance of the LASSO versus the Random Forest models, and includes a post-selection visualization that re-evaluates the exploratory analysis using only the genes selected by the model in Chapter 4. Finally, **Chapter 6** synthesizes the general conclusions of the work and establishes future objectives. The report concludes with the bibliographic references and appendices.

Finally, the report concludes with the bibliographic references and appendices.

1 Data preprocessing

Note: It is important to clarify that the gene expression matrix (RNA-seq) used in this study was already previously normalized.

Performing statistical analyses when the number of variables (genes) vastly exceeds the number of observations (patients) presents enormous mathematical and computational complexity. To mitigate this high dimensionality problem, it is essential to resort to data preprocessing techniques, applying an initial filter that simplifies the dataset without altering the integrity and viability of the results.

1.1 Variance Filtering

In this project, we opted to apply a filter based on the variance criterion. The choice of this method is supported by scientific literature. According to [4], variance filtering stands out as a viable and highly effective approach in genomic data analysis. The removal of genes with low variability significantly increases the statistical power in identifying differentially expressed genes, keeping the False Discovery Rate (FDR) under control and at appropriate levels. The statistical and biological rationale underlying this decision is simple: genes whose expression remains constant (or nearly constant) across all patients possess no discriminatory power to assist in distinguishing between oligodendroglioma and astrocytoma.

Guided by this principle, the filtering process was implemented by establishing the 10th percentile as the cutoff threshold. Consequently, only the genes whose variance was strictly greater than this value were retained in the dataset.

Through the application of this elimination technique, the original dataset underwent a considerable noise reduction, decreasing from the initial 20 501 variables to a total of 18 451

genes with greater biological and mathematical potential for the subsequent exploration and modeling phases.

1.2 Exploratory Analysis

After applying the variance filter, and before advancing to the modeling phase, it is fundamental to understand the structure of the resulting data. The main objective of this phase is to establish a visual comparison point of the behavior of the 18 451 variables, allowing us to contrast these preliminary results with the data clustering after the application of the models (which will be explored in Chapter 5).

To this end, two distinct unsupervised dimensionality reduction techniques were used to map the high-dimensional data into a visually interpretable 2D space: a linear approach (**PCA**) and a non-linear approach (**UMAP**).

PCA is a classical linear technique. As detailed by [7] and [8], PCA allows summarizing a large set of correlated variables into a smaller number of representative variables, called principal components, which are linear combinations of the original variables and are strictly uncorrelated with each other. This mathematical transformation is constructed so that the first few components retain the maximum possible variance information from the original data, offering an excellent perspective of the global structure and identifying the axes of greatest genetic variability.

UMAP, on the other hand, is a more advanced non-linear technique. As explored by [7], this algorithm stands out for its superior ability to model complex mathematical relationships, simultaneously preserving the local structure and the global topology of the data. In biological practice, UMAP can frequently identify patient clusters more clearly and sharply than linear approaches by better handling the non-linear behavior of gene expression.

The resulting projections of these two approaches can be observed in Figure 1 and Figure 2. Analyzing Figure 1, PC1 explains 11.6% of the variance, and PC2 explains 9.8% of the variance, as seen both in the plot and in the summary. Together, PC1 and PC2 explain 21.4% of the variance. By visualizing the plot, the separation between both types of gliomas is clear: astrocytomas lean predominantly toward the upper-left quadrant, while oligodendrogliomas span the lower-right quadrant.

On the other hand, Figure 2 demonstrates the power of the non-linear approach. Even using the dataset solely with the initial variance filter (18 451 genes), UMAP reveals different regional aggregations, although some overlap between the two subtypes is also observed. Astrocytoma samples tightly aggregate at the bottom-right tail of the curve, whereas oligodendroglioma samples heavily dominate the top-left upper bounds.

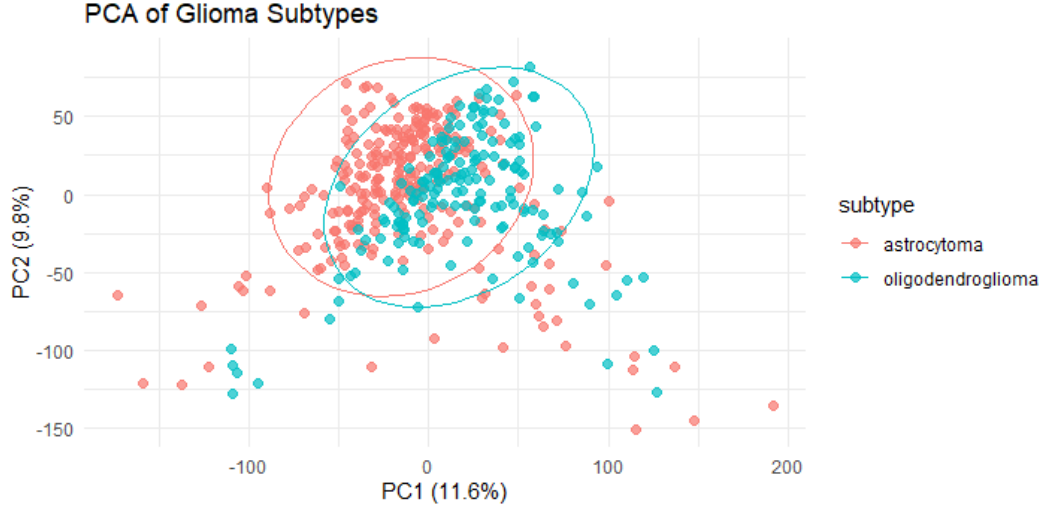


Figure 1: PCA projection using 18 451 genes.

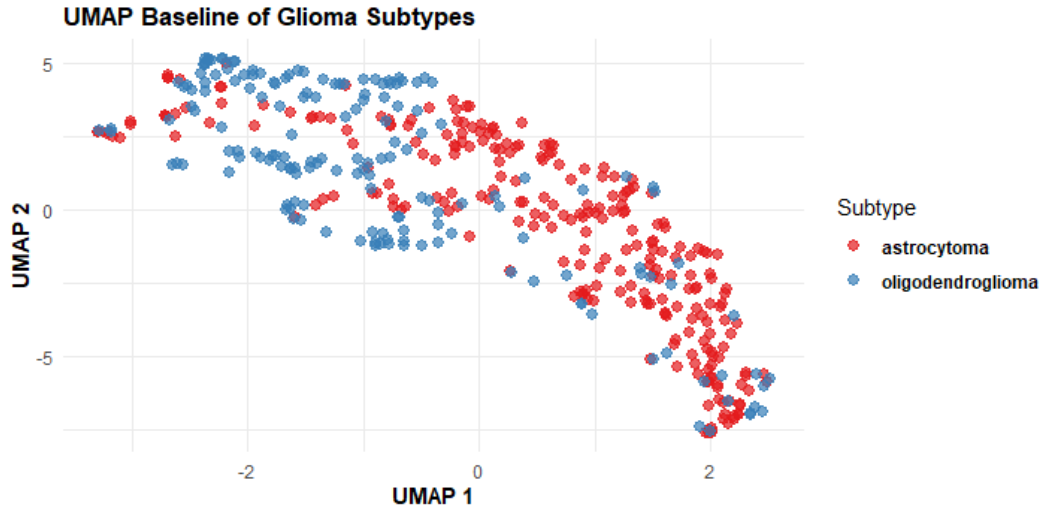


Figure 2: UMAP projection using 18 451 genes.

1.3 Data split

The first step in building a robust predictive model involves adequately training the algorithm and ensuring that its evaluation simulates clinical reality: diagnosing patients the model has never seen. As highlighted by [8], splitting the dataset into independent subsets for training and testing is a critical and foundational step in statistical learning for assessing model performance. To achieve this, the filtered dataset was divided into two groups: the training set and the test set. This methodological separation is strictly necessary to avoid the phenomenon of overfitting. As detailed in the previously cited reference ([8]), overfitting occurs when an algorithm adjusts too closely to the training data. This can happen in two ways: either the model simply "memorizes" the exact answers of the observed data, or it begins to follow and model the random errors and statistical noise inherent to that partic-

ular sample. If a model is trained and evaluated with the same data, its apparent efficacy will be illusory, as the modeled noise will not apply to new observations. The central goal of machine learning is not memorization, but rather learning the true underlying genetic patterns to correctly classify new cases of oligodendroglioma and astrocytoma.

In this project, a 75/25 split was chosen. Specifically, 75% of the data ($n = 305$ patients) were allocated to the training set, where the models will optimize their parameters and learn to distinguish the tumors. The remaining 25% ($n = 105$ patients) were isolated and locked away during the entire training phase, serving exclusively as the final test set to validate the real efficacy of the predictions.

2 Random Forest Method

2.1 Fundamentals of Decision Trees

Random Forest is a machine learning model whose fundamental unit is the decision tree. In statistical learning, there are two main types of trees: regression trees and classification trees. Since our problem aims to determine which subtype of lower-grade glioma a patient belongs to, a binary scenario, the use of classification trees was the mathematically appropriate choice. As described in [1] and [8], the structure of a classification tree contains a series of logical splits. The algorithm starts at a root node, which holds the entire sample, and divides the data into internal nodes. Each internal node represents a test on a specific variable. The data flows through these branches until reaching a leaf node, which determinates the final class prediction. Figure 3, taken from doc[8], provides a generic visual example of this structure, showing how data is filtered step-by-step through decision rules until it reaches a final classification.

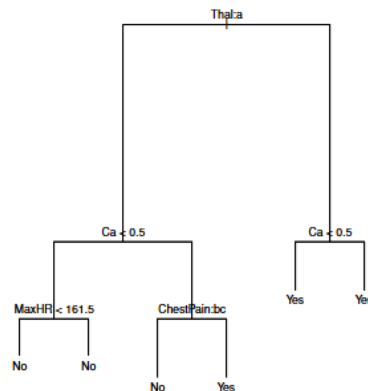


Figure 3: Illustrative example of a classification tree (ref: [8]).

2.2 Model Configuration and Splitting Criteria

While a single decision tree is highly interpretable, it is frequently susceptible to overfitting. To overcome this limitation, Random Forest generates a large number of independent decision trees during training and merges their results (extracting the majority class to determine the final prediction). As elucidated in [8] and [1], this aggregation of multiple "weak" models drastically reduces variance. The model was implemented in R with two hyperparameters governing its architecture:

- **Number of trees (ntree = 500):** Building 500 decision trees ensures that error metrics, such as the OOB error, stabilize, ensuring consistent results.
- **Number of variables randomly sampled as candidates at each split (mtry = 135):** At each branch of the internal nodes, the algorithm evaluates only a random subset of genes to decide the best cut. Relying on the literature [8], the optimal value for classification usually corresponds to the square root of the total number of predictors ($\sqrt{18451} \approx 135.8$). This restriction forces the trees to explore different regions of the genetic space, ensuring their decorrelation.

2.3 Error Metrics and Purity Measures

The Random Forest provides two internal error estimates that offer an initial assessment of predictive performance without requiring a separate validation set:

- **Out-of-Bag (OOB) Error:** Each tree is trained using a random sample with replacement (bootstrap) of the original data. Consequently, a part of the training data is left out of the construction of each individual tree. These excluded observations are called Out-of-Bag. The model tests each tree using its respective OOB data, providing an unbiased and highly reliable estimate of the model's prediction error on unseen data.
- **Class Error:** In addition to the global error, the model calculates the specific error rate for each category. This allows us to evaluate if the algorithm has a bias toward incorrectly classifying a specific subtype, for instance, if it struggles more to diagnose astrocytomas than oligodendrogliomas.

These metrics provide a reliable global picture of the model's performance. Nonetheless, they do not capture the quality of the individual partitions formed at each node. As mentioned in [8], two complementary purity measures are commonly employed for this purpose: the Gini Index and the Entropy.

The Gini Index is defined by Equation (1):

$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk}), \quad (1)$$

where $\hat{p}_{mk}$ denotes the proportion of training observations belonging to class k . A node is considered pure when all observations belong to a single class, driving G toward zero.

The complementary measure, Entropy, is given by Equation (2):

$$D = - \sum_{k=1}^K \hat{p}_{mk} \log(\hat{p}_{mk}). \quad (2)$$

Equation (2) carries an analogous interpretation: since $0 \leq \hat{p}_{mk} \leq 1$, entropy approaches zero as nodes become increasingly homogeneous. Both criteria are numerically similar in practice and considerably more sensitive to node purity than the raw misclassification rate, making them preferable for evaluating split quality.

It should be noted, however, that neither measure was explicitly computed in the present implementation. The `randomForest` package in R handles node splitting internally, and the OOB error and class error reported above were the primary metrics used to evaluate the model’s performance.

2.4 Confusion Matrices

The algorithm’s performance was visually evaluated through Confusion Matrices. Table 1 presents the matrix resulting from evaluating the model on the training data, where an almost perfect performance is naturally expected (only a few misclassifications for both glioma subtypes). Table 2 presents the matrix generated by the predictions on the test data, providing the real metric of the model’s diagnostic accuracy on unseen data. We observe that there is 1 misclassification: a patient with Oligodendroglioma is incorrectly diagnosed by the model with Astrocytoma.

Table 1: Training-set confusion matrix for the Random Forest model.

True	Predicted		
	Astrocytoma	Oligodendroglioma	class.error
Astrocytoma	178	4	0.02197802
Oligodendroglioma	2	131	0.01503759

Table 2: Test-set confusion matrix for the Random Forest model.

True	Predicted	
	Astrocytoma	Oligodendroglioma
Astrocytoma	72	0
Oligodendroglioma	1	32

2.5 Strengths and Limitations

Like any machine learning approach, applying the Random Forest algorithm to high-dimensional genomic data brings clear advantages, but also some practical limitations.

Strengths:

- **High Accuracy:** The ensemble nature of the model provides a highly predictive capability. By aggregating the decisions of hundreds of trees, individual errors cancel each other out, leading to extremely accurate classifications.
- **Resistance to Overfitting:** Unlike a single decision tree, which often grows until it simply memorizes the training data, the forest introduces randomness through bootstrap sampling and by restricting the variables checked at each split. This dynamic makes the model particularly effective at preventing overfitting.

Limitations:

- **Computational Load:** Building 500 decision trees and evaluating thousands of cut-points across 18 451 genes is a demanding process that requires significant time and processing power.
- **Lack of Direct Gene Selection:** In the context of our project, this is the main drawback. Even though Random Forest can rank genes by Feature Importance, it does not actively discard the irrelevant ones. The final model still relies on almost all the original variables, failing to create a truly sparse model.

In a real clinical environment, the ultimate goal is to identify a restricted set of biomarkers rather than relying on thousands of genes to make a diagnosis. The Random Forest’s inability to strictly isolate the most important variables highlights the need for other models, leading us naturally to Logistic Regression.

3 Logistic Regression Model

Let $X = \{x_1, \dots, x_N\}$ be the set of observations, where each patient x_i has p variables (in this case, gene expression) alongside a binary label $y_i \in \{0, 1\}$. In this work, class 1 represents patients with Oligodendroglioma and class 0 represents patients with Astrocytoma. Logistic regression operates under the assumption that the log-odds of an observation belonging to class 1 can be modeled with a linear regression, represented in Equation (3). According to the formulations in the reference literature [8] and [2], we have:

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i \quad (3)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ corresponds to the coefficients associated with each variable, β_0 is the independent term of the regression, and $\pi_i = \mathbb{P}(Y_i = 1 \mid X_i = \mathbf{x}_i)$, the conditional probability of belonging to class 1. Rearranging this expression yields the sigmoid function, which maps any linear combination of the predictors to a probability value in the interval $[0, 1]$, as defined in Equation (4):

$$\mathbb{P}(Y_i = 1 \mid X_i = \mathbf{x}_i) = \frac{1}{1 + \exp[-(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)]} \quad (4)$$

This function is governed by $p + 1$ parameters and produces a continuous score between 0 and 1 for each observation. The final class assignment follows a decision threshold of 0.5: an observation is predicted as class 1 whenever $\pi_i \geq 0.5$, and as class 0 otherwise.

In traditional logistic regression, the β_0 and β coefficients presented in the previous equation would be estimated by minimizing the negative log-likelihood function. However, our dataset presents a severe discrepancy: the number of predictor variables (genes) vastly exceeds the number of observations (patients), meaning $p \gg n$. Confronted with the curse of dimensionality, classical estimates collapse, preventing the algorithm from converging and generating infinite variance and extreme overfitting. To make the modeling of these probabilities viable and correctly estimate the coefficients, it is mathematically mandatory to minimize a penalized negative log-likelihood function.

4 Sparse Logistic Regression with L1 Regularization

4.1 The Elastic Net and LASSO Approach

As outlined above, the introduction of a penalty term into the likelihood function is a mathematical necessity in high-dimensional settings such as the one at hand. As described in the bibliography [3], the Elastic Net approach combines the properties of the Ridge penalty (L_2) and the LASSO penalty (L_1) into a single algorithm. In the same document, the Elastic Net penalty is presented as (5):

$$(1 - \alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \quad (5)$$

where

$$\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2, \quad (6)$$

$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j|. \quad (7)$$

The parameter α ($0 \leq \alpha \leq 1$) controls the weight of the Ridge component (L_2). However, for the modern computational implemented in this report, is the adaptation presented in [2]. In practice, the β coefficients are estimated by finding the minimum of the following adapted objective function:

$$\min_{\beta_0, \beta} \left\{ \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{p}_i) + \lambda \left[\frac{1 - \alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right] \right\} \quad (8)$$

In Equation (8):

$\frac{1}{n} \sum_{i=1}^n L(y_i, \hat{p}_i)$ represents the loss function (the error between the predicted probability and the patient's actual diagnosis), λ is the tuning parameter that controls the overall strength of the penalty and α is the mixing parameter between the two penalizations. The great power of this formulation in the genomic context is revealed precisely when we set $\alpha = 1$. In this scenario, the Ridge component is completely nullified, and the model assumes a strictly LASSO penalty. This L_1 mathematical constraint actively forces the coefficients of less relevant genes to shrink to exactly zero, acting as a rigorous feature selection mechanism. Consequently, in this project, the chosen option was exclusively the L_1 penalty, meaning we did not work with the L_2 penalization.

4.2 Cross-Validation and Parameter Selection

Unlike the Random Forest model explored in Chapter 2, which does not need an external cross-validation because it calculates its own Out-of-Bag (OOB) error using bootstrap samples, Logistic Regression requires a rigorous external approach to tune the penalty hyperparameter λ . The value of λ dictates the severity of the model: a value that is too low results in overfitting (similar to classical logistic regression), while an excessively high value forces all coefficients to zero, resulting in underfitting. To determine the optimal value of λ , a k -fold Cross-Validation method was applied ([8]). In this project, we used 5 folds ($k = 5$). The training set is partitioned into 5 roughly equal-sized subsets. The algorithm is trained on 4 folds and tested on the remaining fold, repeating the process iteratively until all folds have served as validation data. This methodology calculates the deviance curve, shown in Figure 4, which visually illustrates how the error behaves as we increase the model’s penalization.

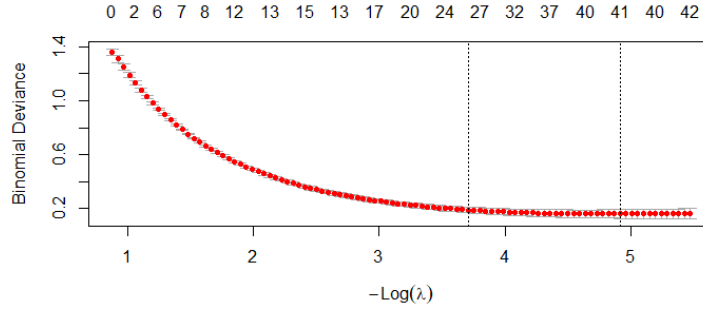


Figure 4: Cross-validation curve used to select the regularization parameter.

From this analysis, we extract two critical reference values, represented in Table 3: λ_{min} , which minimizes the cross-validation error (maximizing predictive accuracy), and λ_{1se} , which represents the most penalized and sparse model within one standard error of the minimum error. Since the goal is to reduce the number of genes, λ_{1se} is selected.

Table 3: Cross-validation results for the selected regularization parameters.

	λ	Index	Measure	Nonzero
min	0.007311	88	0.1614	41
1se	0.024505	62	0.1915	26

4.3 Confusion Matrices

With the penalized coefficients fixed, the performance of the LASSO model was evaluated, similar to the Random Forest, using Confusion Matrices. Table 4 illustrates the confusion matrix corresponding to the training data. Subsequently, Table 5 documents the algorithm’s generalization capability when applied to the test dataset, providing the final metric of its diagnostic precision on unseen samples. We observe that there are 2 misclassifications: True Oligodendroglioma patients were incorrectly diagnosed with the other glioma subtype.

Table 4: Training-set confusion matrix for the Sparse Logistic Regression with L1 Regularization.

True	Predicted	
	Astrocytoma	Oligodendroglioma
Astrocytoma	180	2
Oligodendroglioma	1	132

Table 5: Test-set confusion matrix for the Sparse Logistic Regression with L1 Regularization.

True	Predicted	
	Astrocytoma	Oligodendroglioma
Astrocytoma	72	0
Oligodendroglioma	2	31

4.4 Strengths and Limitations

The application of the L_1 penalty to logistic regression presents remarkable properties, although it carries some mathematical limitations.

Strengths:

- **Simultaneous Feature Selection and Shrinkage:** The main strength of LASSO lies in its ability to execute, in a single algorithmic step, coefficient shrinkage to prevent overfitting and rigorous gene selection. By forcing coefficients to zero, the model generates a highly sparse, clinically applicable, and interpretable genetic signature.
- **Computational Efficiency:** The LASSO convex optimization algorithm is remarkably fast and efficient at handling high-dimensional data.

Limitations:

- **L_1 Norm Limitation in $p \gg n$ Scenarios:** As stated in [3], the L_1 norm has an intrinsic limitation when the number of predictors vastly exceeds the number of observations. LASSO selects at most n variables (in this case, limited to the number of patients in the training set) before the model saturates. Additionally, in the presence of a group of highly correlated genes, LASSO tends to arbitrarily select only one gene from the group, ignoring the rest (a flaw that the Ridge component of the Elastic Net attempts to mitigate).
- **Sensitivity to Outliers:** The traditional penalized logistic model does not correct for deviations caused by outliers in the gene expression data. These atypical values can severely distort the loss function and, consequently, affect the selection of coefficients and variables.

Recognizing this vulnerability highlights that our current methodology still has room for evolution. As a future approach, the implementation of a Robust Classification model would constitute an essential step to accommodate the biological noise inherent to genomic sequencing.

5 Model Comparison and Gene Signature Analysis

Having established and evaluated the Random Forest and LASSO models individually in the previous chapters, this section provides a direct comparative analysis of their performance. Furthermore, it explores the biological relevance of the extracted gene signature through post-selection dimensionality reduction techniques.

5.1 Performance LASSO vs Random Forest

Following the optimization of both models, their efficacy was compared through accuracy and AUC, as presented in Table 6.

Table 6: Predictive performance of the fitted classification methods.

Method	Accuracy	AUC
Random Forest	99.048%	1.00
L1 Regularization	98.095%	1.00

In terms of Accuracy, both algorithms demonstrated exceptional performance in distinguishing between glioma subtypes: the Random Forest model achieved an accuracy of 99.048%, while the logistic regression model with LASSO penalty reached 98.095%. Beyond overall accuracy, the primary metric for evaluating binary classification models is the **AUC**. The ROC (Receiver Operating Characteristic) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied, plotting the True Positive Rate (Sensitivity) against the False Positive Rate (Specificity) ([8]). The AUC value ranges from 0.5 (random guessing) to 1.0 (perfect classification). Notably, the generated plots shown in Figure 5 and Figure 6 revealed an AUC of exactly 1.0 for both models, theoretically indicating an impeccable separation capacity.

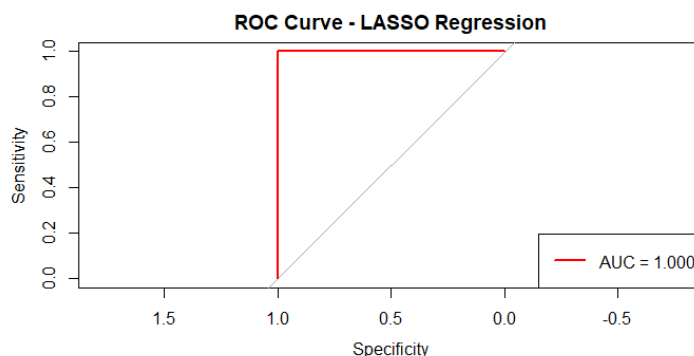


Figure 5: ROC curve and AUC for the LASSO logistic regression model.

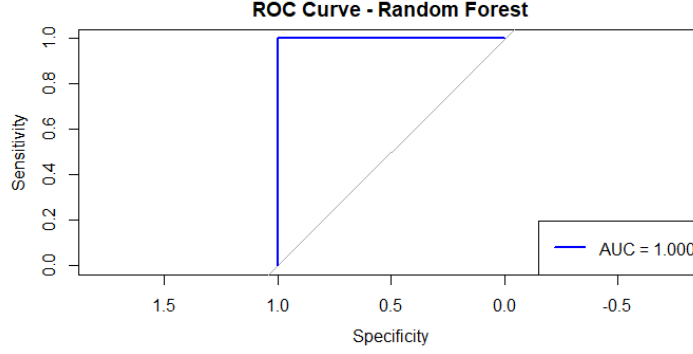


Figure 6: ROC curve and AUC for the Random Forest model.

However, it is important to note that a perfect ROC AUC is compatible with a small number of misclassifications because the ROC AUC assesses ranking across all decision thresholds, whereas the confusion matrix is based on a single classification threshold.

Moving beyond these global performance metrics, the variable extraction provided clinically vital results. Table 7 lists the 26 most crucial genes extracted by the Random Forest (via its Feature Importance criterion) compared to the 26-gene signature retained by the LASSO algorithm. The intersection contains 4 fundamental genes selected simultaneously by both methodologies, shown in 7. The four genes identified form a robust genetic signature. *UBA2* and *LRRC41* are well-established biomarkers in the biological literature specifically associated with lower-grade glioma molecular subtypes ([9]). Crucially, the selection of *TTC4* highlights the biological accuracy of the models, as this gene is located on a chromosome region (1p) whose fractional loss is strongly associated with glioma pathology and subtype differentiation ([10]). Meanwhile, *FBXO42* is a known fundamental regulator of cell cycle and tumor progression ([11]). Their simultaneous selection by two distinct, high-dimensional mathematical algorithms reinforces the biological validity and consistency of the obtained results.

Table 7: Comparison between the genes selected by Random Forest and Sparse Logistic Regression with L1 penalization.

Gene Set	Genes
Genes (Random Forest)	<i>PHACTR4, DRG2, PSMB2, NADK, ASAP3, TTC4, GRPEL2, RPF1, GNAI3, PSMA5, UBA2, BCL10, TXLNA, GTF2B, LRRC41, FBXO42, WDR78, C1orf109, SH3GLB1, GNL2, TRAPPC3, LOC148189, USP48, SCP2, NMNAT1, PHF13</i>
Genes (Sparse Logistic Regression with L1 penalization)	<i>TCL1A, CT45A2, GDAP2, ZNF684, TTC4, TM2D1, FBXO42, BEND5, LRRC41, MMP3, TMEM167B, ZNF181, FNBP1L, RTCD1, FAM54B, LSM14A, GRIK5, LRP3, CAPZB, UBA2, CMPK1, UQCERS1, PABPC4, TMEM97, ELOVL5, ST3GAL3</i>
Genes (Random Forest) $\cap$ Genes (Sparse Logistic Regression with L1 penalization)	<i>TTC4, UBA2, LRRC41, FBXO42</i>

Biomarker Disjunction: Random Forest vs. LASSO regression

Intersection of the 26 selected genes for Glioma classification

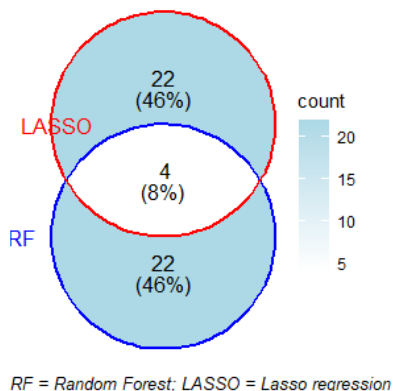


Figure 7: Venn diagram showing the percentage of the genes selected by Random Forest (blue) and Sparse Logistic Regression with L1 penalization (red).

5.2 Post-selection Visualization

To validate the robustness of the isolated genetic signature, the initial exploratory analysis (Chapter 1) was re-evaluated, this time restricting the multidimensional space exclusively to the 26 genes selected by the LASSO model. Figure 8 displays the patient projections resulting from **PCA** and Figure 9 displays the projection obtained with **UMAP**.

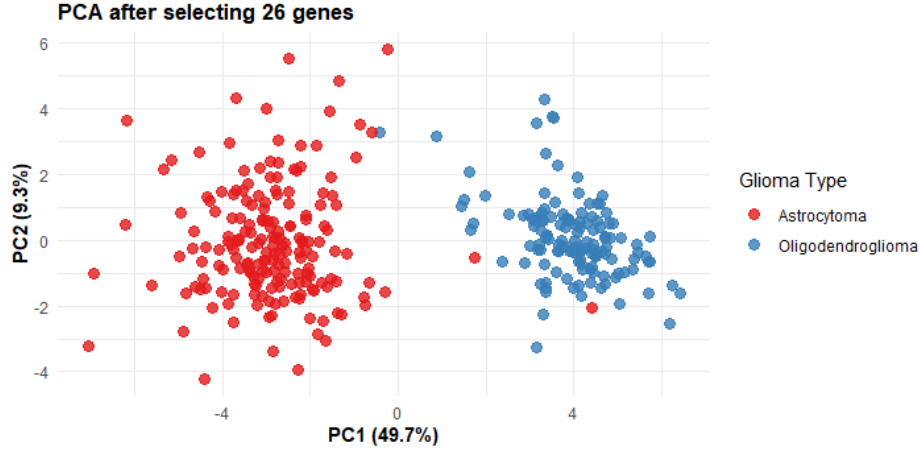


Figure 8: PCA projection using the 26 genes selected by the LASSO model.

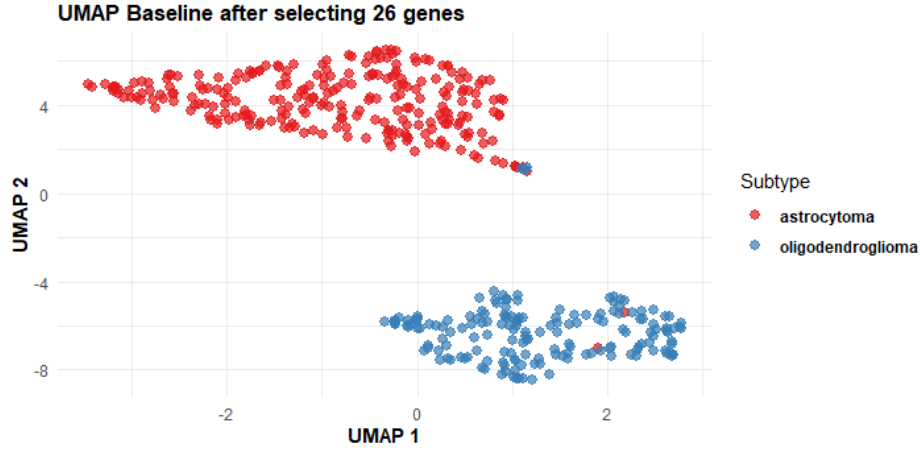


Figure 9: UMAP projection using the 26 genes selected by the LASSO model.

The visual results confirm the high discriminatory power of this reduced signature. In the PCA plot, the first principal component (PC1), responsible for explaining a substantial 49.7% of the data variance, creates a striking linear separation: patients diagnosed with Astrocytoma (in red) cluster predominantly in the negative quadrant of the PC1 axis (left), while Oligodendroglioma cases (in blue) are isolated in the positive quadrant (right). The non-linear UMAP method structurally corroborates this separation. The projection highlights the formation of two highly dense primary clusters: Astrocytoma patients are distinctly aggregated in the upper region of the plot, in clear opposition to Oligodendroglioma patients, who are confined to the lower region.

Nonetheless, it is important to note that in both dimensionality reduction techniques, there are noticeable outliers. The presence of red dots interspersed within the blue dispersion cloud, and vice versa, is visible. These residual overlaps reinforce the conclusion established in the previous chapter: the variability and noise inherent to genomic data are not entirely mitigated by classical LASSO.

6 Conclusion

Within the timeframe of one week, the objectives initially outlined for this project were successfully met. Two classification models were rigorously developed, trained, and evaluated on a high-dimensional RNA-seq dataset comprising astrocytoma and oligodendroglioma patients: a Random Forest method and a Logistic Regression with L1 regularisation (LASSO). Both models demonstrated strong predictive performance, achieving test accuracies of 99.05% and 98.1%, respectively, while simultaneously enabling the identification of a compact and biologically interpretable set of 26 genes. Dimensionality reduction techniques (namely **PCA** and **UMAP**) confirmed the discriminative power of the selected gene signatures, revealing clear separation in the reduced feature space. Nevertheless, certain limitations must be acknowledged.

The **AUC** reached a value of 1.0 for both models. As previously established, this perfect score is mathematically compatible with the minor misclassifications observed in the confusion matrices, confirming the strong discriminatory capacity of the algorithms. Despite these excellent results, the inherent genomic noise and the presence of outliers remain biological realities that must be carefully addressed in future modeling stages.

On the modelling side, the implementation of a Robust Logistic Regression model represents a natural next step, as it is specifically designed to attenuate the leverage of anomalous observations and produce more reliable coefficient estimates in the presence of outliers. This approach has been explored in the statistical literature as a principled solution to improve the resilience of logistic models under data irregularities, and constitutes a natural starting point for the implementation of this method ([5] and [6]).

Overall, this study establishes a highly promising framework for glioma subtype classification. Finally, the identification of this concise genetic signature highlights how important these variables can be for understanding glioma heterogeneity, improving clinical diagnosis, and supporting the development of targeted therapies.

7 References

- [1] Baladram, S. *Random Forest, Explained: A Visual Guide with Code Examples*. Available at: <https://medium.com/data-science/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c>. (2024). Accessed on: June 18, 2026.
- [2] Carrilho, J., & Lopes, M. *Classification and biomarker selection in lower-grade glioma using robust sparse logistic regression applied to RNA-seq data*. Available at: <https://biometria.ufla.br/index.php/BBJ/article/view/634/350>. (2022). Accessed on: June 18, 2026.
- [3] Zou, H., & Hastie, T. *Regularization and Variable Selection Via the Elastic Net*. Available at: <https://doi.org/10.1111/j.1467-9868.2005.00503.x>. (2005). Accessed on: June 18, 2026.
- [4] Hackstadt, A. J., & Hess, A. M. *Filtering for increased power for microarray data analysis*. Available at: <https://doi.org/10.1186/1471-2105-10-11>. (2009). Accessed on: June 18, 2026.
- [5] Rousseeuw, P. J., & Van Driessen, K. *Computing LTS regression for large data sets*. Available at: <https://doi.org/10.1007/s10618-005-0024-4>. (2006). Accessed on: June 18, 2026.
- [6] Segaeert, P., Lopes, M. B., Casimiro, S., Vinga, S., & Rousseeuw, P. J. *Robust identification of target genes and outliers in triple-negative breast cancer data*. Available at: <https://doi.org/10.1177/0962280218794722>. (2018). Accessed on: June 18, 2026.
- [7] Carnot Research Pvt. Ltd. *Understanding Dimensionality Reduction: PCA vs t-SNE vs UMAP vs FIt-SNE vs LargeVis vs Laplacian Eigenmaps*. Available at: Medium article. (2024). Accessed on: July 12, 2026.
- [8] James, G., Witten, D., Hastie, T., & Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*. Available at: <https://www.casact.org/sites/default/files/2022-12/James-G.-et-al.-2nd-edition-Springer-2021.pdf#page=3.08>. (2021). Accessed on: July 12, 2026.
- [9] Elbashir, M. K., Alanazi, A., & Mahmood, M. A. *Decoding Multi-Omics Signatures in Lower-Grade Glioma Using Protein-Protein Interaction-Informed Graph Attention Networks and Ensemble Learning*. Available at: <https://doi.org/10.3390/diagnostics15222894>. (2025). Accessed on: July 20, 2026.
- [10] Tatevossian, R. G. *Molecular genetic analysis of paediatric low-grade astrocytoma*. Doctoral thesis, UCL (University College London). Available at: <https://discovery.ucl.ac.uk/id/eprint/19560/1/19560.pdf>. (2009). Accessed on: July 20, 2026.
- [11] Hoellerbauer, P., Kufeld, M., Arora, S., Mitchell, K., Girard, E. J., Herman, J. A., Olson, J. M., & Paddison, P. J. *FBXO42 activity is required to prevent mitotic arrest, spindle assembly checkpoint activation and lethality in glioblastoma and other cancers*. Available at: <https://doi.org/10.1093/narcan/zcae021>. (2024). Accessed on: July 20, 2026.

Appendices

All materials produced throughout this project are accessible via the following links. The Google Drive (https://drive.google.com/drive/folders/1ar55b2239-hmNeh3IAJhkBARF04q4vgz?usp=drive_link) contains two files: the patient dataset with all associated gene expression variables (LGG.Rdata), and the complete analytical pipeline implemented in R (Glioma-Classification-Pipeline.R). Additionally, the links (https://github.com/Kaguya-hash/help_me_diagnost) and (<https://help-me-diagnost.onrender.com/>) direct to a web application developed as part of this work, which allows the user to train the classification model and obtain a diagnostic prediction for a given patient profile.