Cihan University-Erbil Scientific Journal (CUESJ)

Research Article

An Elastic Net Approach to Logistic Regression for Genetic Selection in High-Dimensional Brain Cancer Data

Nozad H. Mahmood[1], Dler H. Kadir[1]

[1]*Department of Statistics and Information, College of Administration and Economics, Salahaddin University-Erbil, Erbil, Iraq,*

ABSTRACT

The study explores issues related to the treatment of brain cancer caused by the heterogeneous nature of different variants of brain tumors. The objective of this study was to identify essential genes present in multiple types of brain cancer using high-dimensional gene expression data available on the Curated Microarray Database. The study's dataset comprised a total of 130 samples belonging to four subtypes of brain cancer and 16384 gene expression variables. Thus, the penalized Elastic Net method, in conjunction with Multinomial Logistic Regression, was used to cope with the curse of dimensionality problems. Then, accuracy, Kappa statistics, an area under the curve (AUC), and F1-score were utilized to evaluate measures of the model efficiency. Elastic Net proved to be quite effective in the sense of the extensiveness of the variables included in the analysis and successfully restricted gene level further analysis as well as highlighted subtype-specific expression signatures. The model achieved high precision and AUC values, indicating that, in general, the model had a good ability to distinguish all subtypes with some around perfect scores of AUC. Robust parameter estimation was supplemented with cross-validation and other predictive model validation statistical techniques done in R language programming. Thus, these findings suggest that the best model for evaluating large-scale gene expression data of brain cancers is the use of Multinomial Logistic Regression (MLR) with an Elastic Net regularization. There is ample evidence that these selected genes contribute to and serve as targets for therapy, therefore making this study a good starting point for further investigations with respect to understanding their biological role. The corresponding model is also to be applied to test its validity on some other datasets of a quite different nature. This, in turn, may suggest improved diagnostic, prognostic, and therapeutic options for the brain tumor.

Keywords: Brian cancer, elastic net, regularization techniques, gene selection, multinomial logistic model, high-dimensional data

INTRODUCTION

Brain cancer describes a well heterogeneous array of malignancies that arise within the brain tissue and represent a considerable threat to human health around the globe. Such tumors have great variability in pathology, molecular structure, and clinical behavior, leading to variable prognosis and treatment response rates.[1] It is vital to appreciate the multifactorial nature of genetic and environmental determinants of brain tumor development to achieve advancements in cancer patient outcomes.[2]

Recent high-throughput technologies have made it possible to perform extensive profiling of gene expression across brain tumors and construct such high-dimensional datasets that consist of thousands of genes.[3] However, the analysis of such data has great difficulties because of the "curse of dimensionality" where the maximum number of genes (variables) greatly outnumbers the samples.[4] Thus, such a situation can result in overfitting phenomena where models fit to the noise instead of the signal, making models difficult to interpret and generalize.[5] Therefore, effective gene selection approaches are essential for the analysis of high-dimensional genetic data and pinpointing the strongest target spaces responsible for the different subtypes of brain cancers.

Multinomial logistic regression is used for modeling outcome variables that involve a finite number of categories.[6] It is a generalization of the binary logistic regression for problems that involve more than two classes, making it appropriate for the analysis of data on brain cancer.[7] MLR models the relationship of a set of independent variables (gene expression

Corresponding Author:
Nozad H. Mahmood, Department of Statistics and Information, College of Administration and Economics, Salahaddin University-Erbil, Erbil, Iraq. E-mail: nozad.mahmood@sulicihan.edu.krd

Received: November 21, 2024
Accepted: December 20, 2024
Published: January 20, 2025

DOI: 10.24086/cuesj.v9n1y2025.pp14-23

levels) with a dependent variable which is categorical (subtype of brain cancer) by estimating the log-odds for each brain cancer subtype with respect to a certain base category.[8]

Even though MLR seems to be a powerful method, it has problems in high-dimensional areas. Traditional MLR often suffers from overfitting which leads to unstable estimates and poor predictions in large genomic-wide association studies with limited samples.[9] Moreover, collinearity between the predictor variables of which the standard errors will be large, making individual gene effects less identifiable.[10] Such challenges mean that penalized methods have to be employed for better variable selection and to improve the model's robustness.

The Elastic Net is a contemporary and efficient method for regularization and variable selection which improves the shortcomings of the classical MLR in the examination of high-dimensional MLR data sets.[11] It wallows in the L1 (Lasso) and L2 (Ridge) penalties which provide a level of tradeoff between the selection process and shrinkage of the coefficient of the model parameters.[12,13] Where there is an application of the L1 penalty, it is noted to promote the occurrence of sparsity by shrinking some of the coefficients to exactly zero, thereby performing an automatic variable selection process.[14] Where there is an application of the L2 penalty, all the coefficients are shrunken to zero, including the offset, thereby even alleviating multicollinearity and increasing model stability.

The existing relationship or association between the two penalties means that the application of the Elastic Net to high-dimensional genetic data has a lot of advantages. For instance, the method is less frustrating in identifying relevant genes with respect to the presence of several other genes, thereby devising models that are easier to work with and understand.[15] The performance of the Elastic Net is better than that of the Lasso and Ridge regression in most high-dimensional situations.[16,17] Along with its ability to interpret the model, better prediction accuracy increased, thus the method may work for identifying critical genetic markers for subtypes of brain cancer.

The objective of the research is to apply an Elastic Net-regularized MLR framework to identify the key genes which are responsible for the different subtypes of brain cancer using the high dimensional gene expression dataset. It is also intended to make use of the variable selection attributes of the Elastic Net to obtain gene expression patterns, which will assist in the identification of specific subtypes of brain cancer, helping improve its diagnosis and treatment.

Literature Review

Analysis of gene expression has proven to be an important tool in brain cancer studies because it has helped define molecular subtypes, prognostic indicators, and even therapeutic targets.[18] Microarray and RNA sequencing techniques have been employed to characterize the pattern of gene expression in different types and grades of brain tumors which disclose distinct molecular profiles that correlate with the aggressiveness of the tumor features, response to treatment, and survival of patients.[19,20] These studies underscore the usefulness of gene expression data in making sense of the pathology of brain cancer and the development of individualized treatment approaches as well.

The most important problem in this sphere of research is the analysis of data which has a large number of dimensions. As biological research now can title and measure thousands of genes at once, traditional statistical approaches usually fall short because of multicollinearity, overfitting, and poor accuracy.[21] The "curse of dimensionality" relates to the fact that as the number of dimensions increases, the data becomes increasingly sparse, making the identification of meaningful correlations and patterns in the data much more effortful.[5] In genomics scenarios, this can venture to unstable association and prediction in the results.

Various variable selection strategies have been suggested to deal with the high dimensionality issue. Recursive partitioning techniques, for example, stepwise regression methods are easy to implement but may also be overfitted and unstable as they use a greedy search process. Two popular penalized regression techniques among many that can be useful are Lasso regression and Ridge which are used L1 and L2 penalty regression, respectively. Lasso not only does variable selection but also shrinks the estimates of regression coefficients through the introduction of an L1 constraint which drives some coefficients to zero. Ridge regression, on the other hand, employs L2 penalization and does not choose variables but concentrates them near the zero point. The Elastic Net combines the merits of Lasso and Ridge by dealing with issues associated with highly correlated predictors and allowing for a more effective application of variable selection methods.

According to[6] many categorical dependent variables are best analyzed using techniques such as multinomial logistic regression which is one of the more sophisticated statistical techniques. It incorporates the general concepts of logistic regression into cases where an outcome variable consists of three or more categories. In the case of brain cancer, multinomial versus logistic regression would allow us to predict the subtype of a tumor based on the expression of a bunch of genes. The technique calculates the likelihood of a sample being of a specific subtype by estimating the log odds of the subtypes as a linear combination of the predictors which are (genes) in this instance.

The Elastic Net has turned out to be an effective variable selection tool in situations characterized by high dimensionality, particularly in genomics and other biomedical fields.[14] It is particularly appropriate for gene expression data that has such predictors because genes are usually co-regulated and functionally related. The Elastic Net makes use of both L1 and L2 penalties and thus allows coefficient shrinkage to a certain level of being near zero whilst some relevant variables are still able to be selected. Furthermore, the Elastic Net has been proven useful in determining predictive gene signatures of different cancers, among them breast cancer and leukemia.[22] Its use on gene expression data of brain cancers has a great likelihood of helping to uncover significant genes responsible for subtype change and therefore increasing prediction accuracy.

MATERIALS AND METHODS

Acquisition of Data

The dataset pertaining to brain cancer which was used in this analysis originates from the Curated Microarray Database (CuMiDa). The dataset contains gene expression data consisting of 16384 gene variables in 130 samples for four different types of brain cancer with varying numbers of observations as given in Table 1.

Data Preprocessing

The CuMiDa database is useful in the field of machine learning applications for cancer research and provides microarray-based standardized and prepared datasets to harness computational analyses. Every step was performed carefully during the submission of the brain cancer data set for background correction and removal of any inaccurate probes, thus guaranteeing a robust as well as a reliable analysis process.

Study Design

The purpose of the research was to develop a sparse multinomial logistic regression model and test its possible application in classifying the brain cancer subtypes with the help of Elastic Net shrinkage methods to resolve the overfitting issues and cope with a great number of gene expression data independent variables. Several assessment metrics were used for the comprehensive evaluation of how well the models performed such as Accuracy, Kappa statistics, area under the curve (AUC), and F1-score. K-fold cross-validation has also been applied for the purpose of more reliable estimation by eliminating bias or variance which are factors during fitting models which may result from employing only one random sampling split. In addition, R programming was used to develop these models.

Model Development

For a binary logistic model with two classes (0 and 1), the probability of class 1 given the features is modeled as:

$$p(y=1/X)=\frac{e^{\beta_0+\beta^T X}}{1+e^{\beta_0+\beta^T X}}$$

The probability of class 0 is:

$$p(y=0/X)=\frac{1}{1+e^{\beta_0+\beta^T X}}$$

Where y and X are dependent and independent variables, respectively, and β_0 is intercepted $\beta = \beta_1, \beta_2,\ldots)$

Table 1: Types and observations of brain cancer

Types	Number of observations
Ependymoma	46
Glioblastoma	34
Medulloblastoma	26
Pilocytic_astrocytoma	24
Total	130

A multinomial logistic model is a broader form of logistic regression model useful when there is one dependent variable with more than two categories.[23]

The probability for classes k = 1, 2,…, K-1 is:

$$P(y=k/X)=\frac{e^{\beta_{0k}+\beta_k^T X}}{1+\sum_{j=1}^{K-1} e^{\beta_{0j}+\beta_j^T X}}$$

For the K[th] class (as the reference class), the probability is:

$$P(y=K/X)=\frac{1}{1+\sum_{j=1}^{K-1} e^{\beta_{0j}+\beta_j^T X}}$$

These probabilities ensure that the sum over all classes is 1:

$$\sum_{k=1}^{K} P(y_i=k/X_i)=1$$

Combining the above, for each class k (including the reference class K)

$$P(y=k/X)=\begin{cases}\dfrac{e^{\beta_{0k}+\beta_k^T X}}{1+\sum_{j=1}^{K-1} e^{\beta_{0j}+\beta_j^T X}} & ,\ k=1,2,3,\ldots,K-1\\ \dfrac{1}{1+\sum_{j=1}^{K-1} e^{\beta_{0j}+\beta_j^T X}} & ,\ k=K\end{cases}$$

Here, β_{0k} is the intercept for class k, and β_k is the vector of coefficients for independent variables. The likelihood function for the entire dataset, assuming n independent observations, is the product of the individual probabilities:

$$L(\beta)=\prod_{i=1}^{n} P(y_i=k/X_i)$$

It can be written as below:

$$L(\beta)=\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{e^{\beta_{0k}+\beta_k^T X_i}}{\sum_{j=1}^{K} e^{\beta_{0j}+\beta_j^T X_i}}\right)^{1(y_i=k)} \quad (1)$$

The indicator function of $1(y_i=k)=\begin{cases}1 & if\ y_i=k\\ 0 & otherwise\end{cases}$

The log-likelihood function is obtained by taking the logarithm of the likelihood function by taking the natural logarithm of equation (1).[24]

$$\log L(\beta)=\sum_{i=1}^{n}\sum_{k=1}^{K} 1(y_i=k)\log\left(\frac{e^{\beta_{0k}+\beta_k^T X_i}}{\sum_{j=1}^{K} e^{\beta_{0j}+\beta_j^T X_i}}\right) \quad (2)$$

For simplicity equation (2) can be written as below.

$$\log L(\beta)=\sum_{i=1}^{n}\left[\sum_{k=1}^{K} 1(y_i=k)\left(\beta_{0k}+\beta_k^T X_i\right)-\log\left(\sum_{j=1}^{K} e^{\beta_{0j}+\beta_j^T X_i}\right)\right] \quad (3)$$

To estimate the parameters from equation (3), we can maximize $logL(\beta)$ with respect to $(\beta_0,\beta_1,\ldots,\beta_K)$ or minimize the function of $-logL(\beta)$.

Feature Selection

The $logL(\beta)$ function from equation (3) is derived to install penalized regression techniques trying to adapt the logistic multinomial model. The objective of the model is to control the coefficients which may lead to the ratio of penalized regression, especially in situations where the number of observations n is less than the number of features.[25] The penalized multinomial logistic regression model can be stated as:

$$l_{penalty}(\beta) = -\log L(\beta) + penalty\,term$$

$$l_{penalty}(\beta) = -\sum_{i=1}^{n}\left[\sum_{k=1}^{K} 1(y_i = k)\left(\beta_{0k} + \beta_k^T X_i\right) - \log\left(\sum_{j=1}^{K} e^{\beta_{0j}+\beta_j^T X_i}\right)\right] + penalty\,term$$

Elastic net

The Elastic Net penalty adds together both L1-norm and L2-norm penalties that achieve automatic estimation and variable selection of the model simultaneously.[26,27] To maintain the ratio of Lasso and Ridge penalties on the model, we assume that $\alpha \in (0,1)$ then, the Elastic Net penalty can be written in the form as:

$$Penalty_{Elasti\,net} = \lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p}\left|\beta_{kj}\right| + \lambda_2 \sum_{k=1}^{K}\sum_{j=1}^{p}\beta_{kj}^2$$

$\lambda_1 = \lambda\,\alpha$, $\lambda_2 = \lambda\,(1-\alpha)$

If $\alpha = 0$, we will have Ridge regularization.

If $\alpha = 1$, we have Lasso regularization.

If $0 < \alpha < 1$, we have Elastic Net.

The parameters $\lambda > 0$ control the degree of Lasso and Ridge penalties. This parameter is usually estimated through cross-validation, where the aim is to identify the most optimal one that minimizes an objective function on the validation set.

Then, $logL(\beta)$ for a multinomial logistic model with Elastic Net becomes:

$$l_{Elastic}(\beta) = -\sum_{i=1}^{n}\left[\sum_{k=1}^{K} 1(y_i = k)\left(\beta_{0k} + \beta_k^T X_i\right) - \log\left(\sum_{j=1}^{K} e^{\beta_{0j}+\beta_j^T X_i}\right)\right]$$
$$+\lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p}\left|\beta_{kj}\right| + \lambda_2 \sum_{k=1}^{K}\sum_{j=1}^{p}\beta_{kj}^2$$

Model Evaluation

In the context of precision, accuracy, AUC, and F1-score, they are all essential metrics of interest for the performance of classification models. It is worth stating, however, that they have their own benefits and limitations depending on the problem, and most importantly the class imbalance issue.

Accuracy: The accuracy metric is purely how many predictions were correct out of the total cases predicted. It indicates whether positive and negative results were actually positive and negative. Considering the formula and the method, accuracy is easy to grasp and calculate, and it is effective if the class is balanced.[28]

$$Accuracy = \frac{Nmber\ of\ correct\ predictions}{Total\ Number\ of\ Predictions}$$

The F1 score integrates both precision and recall into a single value by taking their harmonic mean rather than the arithmetic mean in which one value can overshadow the other on account of superlative performance. Simply put, it is used to illustrate how well the model performs.[29]

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

AUC: The AUC, where the curve depicts the true positive rate against the false positive rate with regards to the different threshold settings in the receiver operating characteristic (ROC) AUC range values are as follows:

$0 < \text{AUC} < 1$

$AUC = 0.5$ means no discrimination.

$AUC = 1$ indicates perfect discrimination between classes.

It does serve the purpose of assessing the performance of a model, especially in cases where class imbalance exists.

Cross-validation strategy: Cross-validation (K-fold) is the most widespread technique for assessing a model's performance since it allows a reliable estimation with a low level of bias and variance in the evaluation (IYENGAR, 2024).

RESULTS

In this section, the number of genes chosen by all the gene selection techniques for sparsity regularization and performance measures that may be useful will be discussed. Also, the multinomial logistic models will be assessed in terms of their predictive accuracy, whether these models are selected or not. In the model-building process, Figure 1 is a crucial stage. It determines the optimal values of lambda for Elastic Net regularization which in turn improves the tuning of the models during the training set so that they generalize well to new data. This technique is essential in the quest to build optimal and precise models for the classification of brain subtypes from gene expression data. The figure illustrates the results of cross-validation when the Elastic Net method is applied. For Elastic Net, the lambda that would be optimal is 0.00829563. This number will be employed in further analysis for the construction of final models of the method. Figure 1 depicts the relationship between model complexity and tuning parameters. The increase in lambda achieves a greater penalty on model coefficients which may disturb simpler reconciliation of models with higher deviance. The optimal value of lambda therefore is the one that tends to be in the middle of these two minimization processes.

Table 2 demonstrates the performance of the Elastic Net penalty application using the optimum lambda value

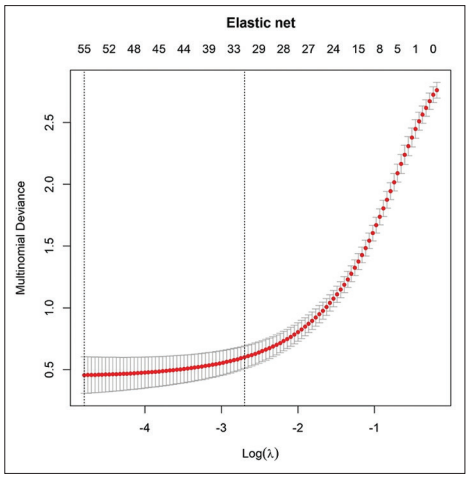

Figure 1: Cross-validation plots to choose the optimal λ to have the minimum deviance

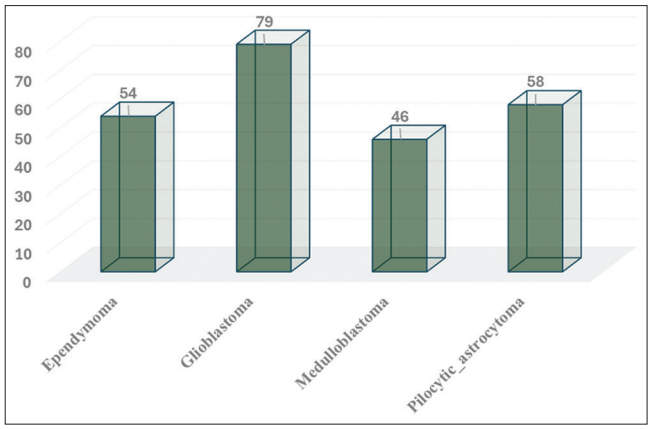

Figure 2: Non-zero coefficients (genes selection) based on classes

established in Figure 1. The performance is assessed according to two metrics, which are Accuracy the proportion of the total samples of each type of brain cancer that were correctly classified, and Kappa an agreement measure concerning the actual and predicted labels taking into account the chance agreement.

The fact that the Kappa value is 0.8294 entails that there is almost perfect agreement between the model outputs and the actual observations. This implies that the model is quite accurate and distinguishes among the several types of brain cancer much better than chance would do. The results show that Elastic Net performs well in accuracy with a score of 87.5 % and a Kappa score of 0.8294. This means that an Elastic Net model with L1 and L2 penalizing factors is suitable for the task at hand both for the fitting of the data and controlling model complexity.

Table 3 shows the number of genes selected (those with non-zero coefficients) by the Elastic Net method for each one of the four brain subtypes. The notable point here is the compression in the dimensionality quite significantly brought about by the Elastic Net penalized method. Elastic Net performs feature selection because some coefficients become zero, leading to sparser models. Sparser models are created by setting specific coefficients equal to zero. As a result, instead of estimating a huge number of 16384 variables, the Elastic Net method has selected 54 variables for Ependymoma, 79 variables for Glioblastoma, 46 variables for Medulloblastoma, and 58 variables for Pilocytic astrocytoma of brain cancers subtypes. Such sparsity may significantly improve both interpretability and computational efficiency. Moreover, the names and coefficients of all the selected genes by Elastic Net are given in Table 1 [Appendix].

Figure 2 shows the number of selected variables (genes) for the Elastic Net method across all four classes. The bar chart clearly demonstrates the sparsity of Elastic Net, highlighting their variable selection capability.

The performance evaluation for the sparsity regularization method of Elastic Net in classifying brain tumor subtypes was conducted and the results are represented in Table 4 and

Table 2: The performance of penalty methods based on the best lambda chosen from cross-validation methods

Criteria	Elastic net
best lambda	0.00829563
Accuracy	0.875
Kappa	0.8294

Table 3: Number of genes selected by Elastic Net regularization method within each class

Classes	Penalty methods
	Elastic net
Ependymoma	54
Glioblastoma	79
Medulloblastoma	46
Pilocytic_astrocytoma	58

Table 4: The values of AUC for each subtype or class

Classes	Area under the curve (AUC)
	Elastic net
Ependymoma	1
Glioblastoma	0.9907
Medulloblastoma	0.9474
Pilocytic_ astrocytoma	0.95

Figure 3. The evaluation metrics used are AUC values, which quantify the overall ability of a model to discriminate a target class among a set of classes (brain cancer subtypes in this case). According to the results, the Elastic Net performs remarkably very well for most subtypes, in some settings the AUC is close to one. In such a way that it consistently attains AUCs of 1 for the Ependymoma subtype, it means it is always perfect in classifying this type of brain cancer. Furthermore, macro-averaged AUC, which is a global measure of performance across all subtypes, also indicates that Elastic Net is the best performer in all with 0.972027.

The effectiveness of the model improves the more the curve approaches the upper left corner. Therefore, all the curves begin with a steep slope. In addition, the graphs of the individual ROC curves of the subtypes also emphasize the performance of Elastic Net further for the subtypes where they obtained perfect AUC.

In general, Elastic Net portrays strong discriminative power in the classification of brain cancer subtypes as indicated by the high AUC values and ROC curves. The results also show that techniques which promote sparsity such as Elastic Net can greatly enhance multinomial logistic regression models, which in turn can be used in classifying brain cancer subtypes with more than a thousand-dimensional gene data.

Table 5 illustrates the performance metric outcomes of the sparsity regularization method of Elastic Net on different subtypes of brain cancers. Generally, among the subtypes of Ependymoma, the Elastic Net performed high on scores (sensitivity, specificity, and F1-score) = 1 which means that those cases were perfectly classified. Moreover, for the other subtypes, it presents an overall good classification achieving a certain acceptance level of sensitivity, specificity, and F1-scores. This means that the results indicate that Elastic Net indeed takes care of the high-dimensional problems in classifying brain cancer. Better classification could be obtained by feature selection which will also lead to more interpretable models. The fact that there were different results within subtypes suggests that one should consider specific characteristics of subtypes before settling on a classification method. Focusing on the genes picked by the algorithms along with their biological relevance would also be important as research may help clarify various mechanisms underlying different types of brain.

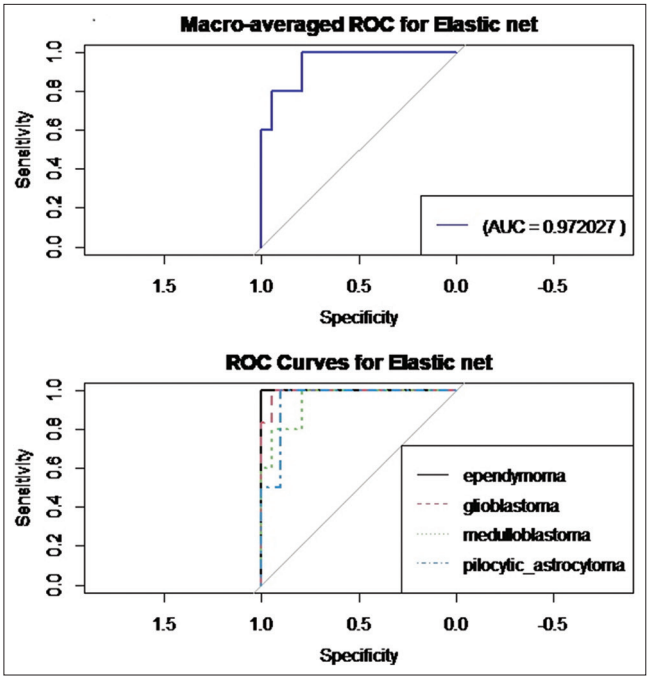

Figure 3: Receiver operating characteristic (ROC) and macro average ROC curves for each method

DISCUSSION

This paper employed an Elastic Net-regularized multinomial logistic regression model to derive gene signatures for the various brain cancer subgroups using high dimensional gene expression data. The results show the usefulness of this method in solving the problems associated with the high dimensionality and multicollinearity of independent variables which leads to reducing the model to be a better performance and more robust interpretation of the model.

The Elastic Net's ability to perform simultaneous variable selection and coefficient shrinkage proved crucial in handling the high-dimensional gene expression data. By shrinking some coefficients to zero, the Elastic Net effectively identified a subset of genes with the strongest association with each brain cancer subtype, reducing the complexity of the model and enhancing its interpretability. This feature selection capability aligns with previous research demonstrating the utility of the Elastic Net in high-dimensional genomic data analysis for various cancer types, including breast cancer and leukemia.[22] Similarly, other studies have emphasized the importance of robust feature selection methods for extracting meaningful insights from complex gene expression data in brain cancer research.[4,20]

The performance evaluation using AUC and other metrics confirmed the strong discriminatory power of the Elastic Net model. Achieving near-perfect AUC values for certain subtypes suggests the potential of this approach for improved diagnosis and subtype classification. This finding echoes the broader trend in brain cancer research where gene expression analysis has been increasingly utilized to identify molecular subtypes and prognostic markers.[18,19] The variability in performance noted between the different observed subtypes reinforces the fact that brain cancer is indeed a complex disease and that there is a need for gene expression signatures specific to each subtype, a view that has been noted in previous studies which have shown that different grades and types of brain tumors possess distinct molecular characteristics.[20]

Table 5: Performance criteria for methods based on the subtypes of brain cancer

Penalty method	Classes	Sensitivity	Specificity	Positive Pred value	Negative Pred value	Precision	Recall	F1-score
Elastic net	Ependymoma	1	1	1	1	1	1	1
	Glioblastoma	0.8333	1	1	0.9474	0.8333	1	0.9091
	Medulloblastoma	0.6	1	1	0.9048	0.6	1	0.75
	Pilocytic astrocytoma	1	0.85	0.5714	1	1	0.5714	0.7273

The biological importance of the genetic variations among subtypes of interest needs to be explored further. Knowing the functions of these genes can help one understand the drivers of each subtype and possibly more importantly, help one find new possible therapies. On the other hand, there are also attempts being made in this area which seek to resolve the combined genetic and environmental epidemiology of brain cancer.[1] Thus, further investigations could examine the incorporation of more clinical and molecular variables with a view to improving the predictive power and clinical applicability of the model. This categorical approach is in line with the shift in opinion toward the role of data integration in precision oncology.[30]

Overall, this study demonstrates that Elastic Net-regularized multinomial logistic regression can be an efficient method for analyzing high-dimensional gene expression data relevant to brain tumor research. The identified subtype-specific gene expression signatures could help advance diagnosis, prognosis, and treatment strategies leading towards better tailor-made approaches to fight against this dreadful disease.

CONCLUSION

The research proceeded with the construction of a sparse multinomial logistic regression model based on Elastic Net shrinkage methods in an effort to classify brain cancer subtypes based on genetic expression data. This solved the overfitting problem and the problem of having too many independent variables which is a common scenario in high-dimensional data problems. The Elastic Net regularization technique was therefore found to be of importance in the analysis of high-dimensional gene expression data by allowing for variable selection and shrinkage of the coefficients at the same time. This enabled the model to select only those groups of genes with the highest significance to risk for a particular brain cancer subtype, thus improving the performance and the explanatory power of the model.

The results confirmed the validity of the Elastic Net-regularized multinomial logistic regression as a competent model for the analysis of high-dimensional gene expression data. This method contributed not only to increasing the accuracy of the classification of subtypes of brain cancers but also enabled the discovery of subtype-shifting gene expression signatures. The discovered gene expression signatures have the potential to provide the needed impetus toward better diagnosis, prognosis, and treatment measures for brain cancer, making it possible to employ more directed and effective measures against this problem. More studies are suggested to determine the biological relevance of the genes selected for this study as well as apply the model to more varied populations for better substantiation.

REFERENCES

1. D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. Von Deimling and D. W. Ellison. The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology*, vol. 23, no. 8, pp. 1231-1251, 2021.
2. Q. T. Ostrom, N. Patil, G. Cioffi, K. Waite, C. Kruchko and J. S. Barnholtz-Sloan. CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2013-2017. *Neuro-Oncology*, vol. 22, no. Suppl 1, pp. iv1-iv96, 2020.
3. Y. Ma and Z. Xi. Integrated analysis of multiomics data identified molecular subtypes and oxidative Stress-Related prognostic biomarkers in Glioblastoma multiforme. *Oxidative Medicine and Cellular Longevity*, vol. 2022, pp. 1-15, 2022.
4. M. Ahdesmäki and K. Strimmer. Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 503-5192010.
5. T. Hastie, T. Robert and J. Friedman. *The elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics*. 2nded., Vol. 10. Springer, Germany; 2009. p. 0387848576.
6. D. W. Hosmer Jr., S. Lemeshow and R. X. Sturdivant. *Applied Logistic Regression*. Wiley, United States; 2013.
7. Pearson Deutschland. *Econometric Analysis. Pearson eLibrary*; 2019. Available from: https://elibrary.pearson.de/book/99.150005/9781292231150 [Last accessed on 2024 Jun 10].
8. N. Mahmood, R. Yahya and S. Aziz. Apply binary logistic regression model to recognize the risk factors of diabetes through measuring glycated hemoglobin levels. *CUESJ*, vol. 6, no. 1, pp. 7-11, 2022.
9. P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data*. Springer, Germany, 2011.
10. K. P. Vatcheva, M. Lee, J. B. McCormick and M. H. Rahbar. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology* (*Sunnyvale*), vol. 6, no. 2, p. 227, 2016.
11. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
12. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, vol. 58, no. 1, pp. 267-288, 1996.
13. N. H. Mahmood, D. H. Kadir, R. O. Yahya and H. Q. Birdawod. The significance of delivery methods and fetal gender in reducing stillbirth rate: Using the generalized regression model. *Clinical Epidemiology and Global Health*, vol. 29, p. 101710, 2024.
14. J. Friedman, T. Hastie and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, vol. 33, no. 1, pp. 1-22, 2010.
15. J. O. Ogutu, T. Schulz-Streeck and H. P. Piepho. Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proceedings,* vol. 6, no. S2, p. S10, 2012.
16. N. Mahmood. *Sparse Ridge Fusion for Linear Regression. STARS*, 2013. Available from: https://stars.library.ucf.edu/etd/2767 [Last accessed on 2024 Jul 03].
17. T. Hastie, R. Tibshirani and M. Wainwright. *Statistical Learning with Sparsity*. CRC Press, United States, 2015.
18. M. Ceccarelli, F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, T. S. Sabedot, B. A. Murray, O. Morozova,… & Y. Newton. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, vol. 164, no. 3, pp. 550-563, 2016.
19. C. Neftel, J. Laffy, M. G. Filbin, T. Hara, M. E. Shore, G. J. Rahme, A. R. Richman, M. E. Shoreet and G. J. Rahmeal. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*, vol. 178, no. 4, pp. 835-849.e21, 2019.
20. Y. Zhang, P. K. S. Ng, M. Kucherlapati, F. Chen, T. Liu, Y. H. Tsang, G. De Velasco, K. J. Jeong and R. Akbani. A pan-cancer proteogenomic atlas of PI3K/AKT/MTOR pathway alterations. *Cancer Cell*, vol. 31, no. 6, pp. 820-832.e3, 2017.

21. G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, Germany, 2021.
22. R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91-108, 2005.
23. J. A. M. Pérez and P. S. P. Martín. Regresión logística. *Medicina De Familia Semergen*, vol. 50, no. 1, p. 102086, 2023.
24. N. H. Mahmood, S. H. Murad and K. K. Kakamad. Ordinal logistic regression for students academic performance in Kurdistan region of Iraq. *Information Management and Business Review*, vol. 10, no. 2, pp. 17-22, 2018.
25. J. E. Yoo. *Penalized Regression in Large-Scale Data Analysis*. Springer, Singapore, pp. 71-91, 2024.
26. C. Wang, N. Li, H. Diao and L. Lu. Variable selection through adaptive elastic net for proportional odds model. *Japanese Journal of Statistics and Data Science*, vol. 7, no. 1, pp. 203-221, 2024.
27. L. Liu, J. Gao, G. Beasley and S. H. Jung. LASSO and elastic net tend to over-select features. *Mathematics*, vol. 11, no. 17, p. 3738, 2023.
28. J. Balayla. *Prevalence Threshold and bounds in the Accuracy of Binary Classification Systems*. Cornell University, New York, 2021.
29. P. Christen, D. J. Hand and N. Kirielle. A review of the F-measure: Its history, properties, criticism, and alternatives. *ACM Computing Surveys*, vol. 56, no. 3, pp. 1-24, 2023.
30. M. Weller, W. Wick, K. Aldape, M. Brada, M. Berger, S. M. Pfister, R. Nishikawa, M. Rosenthal, P. Y. Wen, R. Stupp and G. Reifenberger. Glioma. *Nature Reviews Disease Primers*, vol. 1, no. 1, p. 15017, 2015.

APPENDIX

Appendix: Selected variables (gens) based on each brain cancer subtypes

Ependymoma		Glioblastoma		Medulloblastoma		Pilocytic_astrocytoma	
(Intercept)	**0.82307**	**(Intercept)**	**0.20406**	**(Intercept)**	**−0.43924**	**(Intercept)**	**−0.58789**
X1552296_at	0.02868	X1438_at	0.0189	X1553157_at	0.04591	X1552302_at	0.09649
X1552970_s_at	0.0541	X1552619_a_at	0.09434	X1553268_at	0.04684	X1552304_at	−0.02193
X1553140_at	0.07397	X1553043_a_at	0.03575	X1553611_s_at	0.19474	X1554095_at	−0.02047
X1553321_a_at	0.00582	X1553276_at	0.10793	X1554098_at	0.06079	X1554170_a_at	−0.10002
X1553513_at	−0.06128	X1553411_s_at	0.24225	X1555154_a_at	−0.01896	X1554272_at	−0.01436
X1553622_a_at	0.08222	X1553441_at	0.00096	X1556092_s_at	0.0759	X1555250_a_at	0.06405
X1553734_at	0.09392	X1553983_at	0.00948	X1556444_a_at	0.2522	X1555403_a_at	0.00704
X1554116_s_at	−0.06526	X1553984_s_at	0.07494	X1556963_at	0.01203	X1555836_at	−0.02157
X1555007_s_at	0.14043	X1554112_a_at	−0.04166	X1559735_at	0.00161	X1556619_at	0.15262
X1555082_a_at	0.03328	X1554332_a_at	0.05281	X1560091_a_at	0.18482	X1556776_a_at	−0.04423
X1557636_a_at	0.06628	X1555403_a_at	−0.00685	X1560108_at	0.05542	X1556877_at	0.26185
X1560503_a_at	−0.07319	X1555925_at	0.07318	X1560734_at	0.00653	X1558105_a_at	0.06
X1560834_a_at	0.11466	X1555980_a_at	−0.10482	X1561341_at	0.03356	X1558280_s_at	0.09876
X1561106_at	0.03096	X1556019_at	−0.10273	X1562309_s_at	0.09042	X1559883_s_at	0.00047
X1561429_a_at	0.0762	X1556573_s_at	0.06695	X1569188_s_at	0.04031	X1560647_at	0.03261
X1561504_s_at	0.10127	X1556896_at	0.01523	X1569290_s_at	−0.13844	X1561324_at	0.05663
X1562371_s_at	0.03812	X1557167_at	−0.10269	X1569469_a_at	0.06698	X1561343_a_at	−0.00641
X1562583_s_at	0.01223	X1557169_x_at	−0.03906	X201015_s_at	0.05989	X1563145_at	0.09144
X1563814_at	0.09818	X1557309_at	0.15031	X201418_s_at	0.01648	X1568752_s_at	−0.03865
X1568606_at	0.00813	X1557325_at	0.05706	X201534_s_at	−0.03106	X1568888_at	−0.0945
X1568644_at	−0.0022	X1557359_at	0.28137	X201565_s_at	−0.11082	X1569934_at	0.06197
X1569186_at	0.14491	X1557486_at	−0.01431	X201566_x_at	−0.08521	X200790_at	−0.16455
X1570035_at	−0.00492	X1558508_a_at	0.06359	X201876_at	−0.17757	X201047_x_at	0.18132
X1570169_at	−0.01967	X1558568_a_at	−0.06444	X202381_at	−0.00876	X201133_s_at	0.12042
X200811_at	0.00424	X1558834_s_at	0.03118	X202800_at	−0.02537	X201292_at	−0.02405
X201365_at	0.06873	X1558841_at	−0.02931	X202848_s_at	0.14732	X201303_at	−0.05069
X201564_s_at	−0.15413	X1559111_a_at	0.05312	X203795_s_at	0.02413	X201614_s_at	−0.15487
X202357_s_at	0.05722	X1559462_at	0.01933	X204045_at	−0.02328	X201820_at	0.02687
X202409_at	0.10286	X1559992_a_at	0.00583	X204060_s_at	0.03691	X202330_s_at	−0.191
X202713_s_at	0.02084	X1564936_at	0.11029	X204061_at	0.02919	X202421_at	−0.0564
X203337_x_at	−0.00974	X1564964_at	0.06805	X204088_at	−0.17597	X202503_s_at	−0.02291
X204288_s_at	0.00404	X1565598_at	0.11856	X204130_at	0.11738	X202593_s_at	0.03838
X204298_s_at	0.03724	X1569183_a_at	0.10018	X204465_s_at	0.03012	X202763_at	−0.04589
X204741_at	−0.09191	X1569934_at	−0.03686	X204469_at	−0.17362	X202768_at	0.04172
X204800_s_at	0.10079	X1570255_s_at	0.02612	X204781_s_at	−0.00759	X203022_at	−0.03811
X204877_s_at	0.01762	X200811_at	−0.03401	X205311_at	0.05355	X203213_at	−0.15018
X204932_at	0.30163	X200940_s_at	−0.04208	X205445_at	0.06772	X203358_s_at	−0.05451
X204933_s_at	0.30847	X201193_at	0.00945	X205857_at	0.17959	X203477_at	0.01034
X205161_s_at	0.11716	X201584_s_at	0.05808	X205970_at	−0.04295	X203649_s_at	0.06895
X205186_at	0.02606	X201787_at	−0.0496	X206230_at	0.20859	X203768_s_at	0.00534
X205328_at	0.03462	X201850_at	0.02903	X206271_at	−0.01822	X203839_s_at	0.14987

(*Contd...*)

Appendix: (*Continued*)

Ependymoma		**Glioblastoma**		**Medulloblastoma**		**Pilocytic_astrocytoma**	
(Intercept)	**0.82307**	**(Intercept)**	**0.20406**	**(Intercept)**	**−0.43924**	**(Intercept)**	**−0.58789**
X205421_at	0.14693	X202246_s_at	0.10458	X206403_at	0.03297	X203958_s_at	−0.01045
X205562_at	0.00752	X202409_at	−0.06603	X206422_at	0.16859	X204092_s_at	−0.00804
X205666_at	0.06251	X202580_x_at	0.04053	X206722_s_at	0.03394	X204146_at	−0.01616
X205906_at	0.00174	X202718_at	0.01873	X206893_at	−0.0091	X204156_at	0.05223
X205924_at	−0.00405	X202994_s_at	−0.03581			X204338_s_at	0.0807
X206170_at	0.00279	X202995_s_at	−0.12111			X204339_s_at	0.03569
X206308_at	0.0446	X203362_s_at	0.04662			X204414_at	0.03277
X206335_at	0.03515	X203380_x_at	−0.01402			X204444_at	−0.04664
X206336_at	0.00461	X203865_s_at	−0.05842			X204466_s_at	0.03622
X206683_at	0.08253	X203963_at	0.00813			X204632_at	0.00397
X206773_at	0.09108	X204062_s_at	−0.01086			X205046_at	−0.03811
		X204146_at	0.10208			X205339_at	−0.01853
		X204170_s_at	0.01662			X205466_s_at	0.0428
		X204204_at	0.01318			X205508_at	0.00055
		X204466_s_at	−0.07174			X206729_at	0.14615
		X204508_s_at	0.04399				
		X204600_at	0.05308				
		X204639_at	0.11363				
		X204825_at	0.12085				
		X204877_s_at	−0.02019				
		X204882_at	0.13267				
		X205194_at	0.01384				
		X205306_x_at	0.21862				
		X205542_at	0.11852				
		X205762_s_at	0.0316				
		X205775_at	−0.13669				
		X205931_s_at	0.09853				
		X206018_at	0.03698				
		X206025_s_at	0.00091				
		X206026_s_at	0.07345				
		X206159_at	−0.08246				
		X206172_at	0.02561				
		X206178_at	0.15791				
		X206201_s_at	0.06273				
		X206299_at	−0.15668				
		X206364_at	0.03199				
		X206516_at	0.06593				