Evaluation of machine learning algorithms for classification of Glioma based on gene expression

Aditi Sarathy
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
aditisarathy@cmu.edu

Arnav Gupta
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
arnavg3@andrew.cmu.edu

Ketaki Prakash Ghatole
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
kghatole@andrew.cmu.edu

Mirudhula Mukundan
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
mirudhum@andrew.cmu.edu

Abstract

Background: Glioma is a common malignancy of the brain. It is the leading cause of death due to inefficient tumor classification methods and poor prognosis. We propose classification of subtypes of gliomas using gene expression data from a machine learning approach

Methods: We evaluated four different supervised classification models comprising Naive Bayes, K-nearest neighbor, Logistic Regression and Support Vector Machines. The models were trained to classify between normal and glioma samples and further classify the subtypes of glioma. Five fold cross validation was used to evaluate our models for extracted gene expression data.

Results: The Support Vector Machine algorithm was able to perform the best for given samples in all our classification problems. This was followed by K-nearest neighbor and Logistic Regression bearing comparable results. Naive Bayes showed the poorest results when compared to the other classifiers.

Conclusions: Glioma was successfully classified using all four utilized classifiers. An insight into use of different feature selection/extraction techniques and obtaining the "most informative" genes from these techniques may give us new leads into driver genes for the cancer.

1 Introduction

1.1 Problem

Glioma is the most common cancer of the central nervous system and a leading cause of death due to poor prognosis. The treatment of gliomas greatly depends on accurate tumor classification. The mostly widely used WHO classification classifies gliomas into Astrocytoma, Oligodendroglioma, and Glioblastoma based on their cells of origin [1]. Astrocytoma and oligodendroglioma are low grade while glioblastomas are high grade gliomas. Classification of gliomas and normal cells can lead to early diagnosis of the cancer. Additionally, classifying gliomas to distinguish between the lethal high grade glioblastoma and the low grade gliomas is an unmet need and incorrect classification can greatly affect the treatment plan and its outcome.

1.2 Motivation

Most studies on glioma is driven with an aim to find targets for drug delivery in order to alleviate cancer growth without due regard to the genetic nature of glioma. Previous literature, which are acutely oriented to studies on glioblastoma, mainly comprises wet-lab analysis that include elaborate mass spectrometry methods and exhaustive in vitro and in vivo methods in regulated tumor microenvironments, on a variety of druggable targets [2-4]. Classification of gliomas in clinical settings are mostly based on a priori assumptions based on the histological information [5]. Up until now, such a loose classification method was what drove therapy. The few research groups that did come up with ways that can help in classification either have outdated data with an obsolete way of categorizing glioma [6] or use ML approaches on image data from MRS (Magnetic Resonance Spectroscopy) [7], the data for which is difficult to obtain since it not only requires conducting convoluted MRI experiments but also requires human subjects with different types of glioma. Such subjects are rather difficult to find and it is even more difficult to get their consent for conducting such analysis. There are some studies that use ML classifiers on transcriptome data of glioblastoma to identify inactivation of NF1 in cancerous cells [8]. For this reason, our goal is targeted at simplifying such tedious pre-clinical analysis by using ML methods on gene expression data.

Novel advances in genomic research in combination with machine learning methods have enabled accurate classification of tumors in clinical management of cancers. A key research done on classification of gliomas using microarray gene expression data was a starter analysis of sorts that classify glioma as high grade and low grade, using data sets from Gene Expression Omnibus (GEO) [9]. In our project, we take an extra step to classify gliomas according to their cellular origin.

1.3 Overview of Methods

This project aims to classify between the normal and glioma samples and glioma subtypes using supervised machine learning models. Gaussian Naive Bayes (GNB), k-Nearest Neighbour (k-NN), Logistic regression (LR), and Support Vector Machine (SVM) are used for binary classification as well as multi-class classification. The models for GNB, k-NN and LR were implemented from scratch for the project. SVM was implemented using scikit-learn.

GNB is a probabilistic machine learning model that assumes conditional independence between the genes. k-NN is a non-parametric model that assigns weights to the contribution of each gene based on its distance in the feature space. LR uses a sigmoid function to identify the disease type and tuning the regularization parameters prevents overfitting. SVM utilizes the mutual information between the genes and the class label to train the classifier by finding the best separation hyperplane.

This project compares the performance of all four models for given data using a 5-fold Cross Validation.

1.4 Datasets Analyzed

This project used publicly available data for glioma from The Cancer Genome Atlas (TCGA). Gene expression data in TPM format and clinical information were downloaded. The dataset included 704 samples with around 60,000 features with seven class types. We merged and removed some classes to give three class-types for our model. We also selected the top 300 genes that are known to have a role in brain tumors to improve the quality of our classification. The gene expression data for normal control samples with gene expression data of individuals not diagnosed with glioma, were obtained from GTEx in TPM format. The controls obtained had around 1400 samples, and these were undersampled according to the classification study, in order to address the class imbalance.

2 Methods

2.1 Data Preprocessing

Glioma gene expression data was downloaded using the *gdc-client* on TCGA. The samples belonged to three different projects and therefore we chose to use TPM counts. TPM counts are normalized for a sample and hence can be used as a comparison between projects, which may have different protocols for producing reads. The data downloaded was mapped STAR counts for genes and was in separate

files for each sample. These were merged together and only the TPM values were chosen. Labels were created using the provided clincal data. Tissue-matched normal controls were downloaded from GTEx portal, and were avaliable as $\log_2(TPM + 0.001)$, therefore the cancer dataset was also transformed to $\log_2$ values. After this we only selected the top 300 genes reported to be heavily mutated in gliomas. Features that had zero counts over all samples were removed. Also, those who had zero counts in a particular project were also removed. The remaining features were scaled to zero mean and unit variance for final classification. At the end we were left with 575 samples and 90 features

2.2 Gaussian Naive Bayes

Starting with understanding Naive Bayes, it is a generative model that associates a particular finite, discrete label to some problem description. On top of this, in this model, we assume that the features that describe this problem are not dependent on each other. For example, if we want to observe how a particular phenomenon P is influenced by gene G1 and gene G2, then even if the expression of G1 influences G2 or vice-versa, we consider each of the expression as independently contributing to the probability of P occurring. Backed by the Bayes theorem, the posterior probability of an event being labeled as belonging to a particular class is given as: $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$, where $P(c|x)$ is the posterior probability, $P(x|c)$ is the likelihood of the event occurring if it belonged to that particular class label c, $P(c)$ is prior probability for the class label, and $P(x)$ is the prior of the predictor. Calculating likelihood for each combination of features as $P(x_1, x_2, x_3, .., x_n|c)$ is difficult. Even if we have 2 possible outcomes of each feature x, then we would have 2_n parameters in our joint distribution table that will need calculation. Applying our naive assumption here changes the form as $P(x_1, x_2, x_3, .., x_n|c) = P(x_1|c)P(x_2|c)..P(x_n|c)$, which drastically reduces the number of parameters to 4n now, which is much more computation-friendly. This is the rationale behind why we assume this sort of conditional independence between features and the reason behind why we call it "naive".

The above formalism assumes discrete-valued or categorical feature vectors of samples, that make use of a Bernoulli distribution to estimate the likelihood. Now, if we want to extend this logic to features that have continuous values, we simply exchange the likelihood function to a Gaussian distribution, by making the assumption that each feature conforms to a normal distribution, having no co-variance between features themselves. Each feature for every class has a distinct mean and variance that is computed as the parameters of the training data. This is made use of to evaluate the probability of a test variable falling into the Gaussian characteristics of that particular feature, following which an MLE or MAP estimation can be performed depending on the likelihoods obtained for each class. In our project, we define the expression of each gene as a feature, and the individuals or patients as samples that make up the feature vector space. These form the inputs in this project's self-implemented Gaussian Naive Bayes model.

2.3 K-Nearest Neighbour

K-Nearest Neighbour (k-NN) is a supervised machine learning algorithm used for classification. It is a very widely used classification method for real life scenarios as it is a non-parametric method and makes no prior assumption about the data, it is easy to implement and is a very efficient classifier. In k-NN a new test data is classified as the majority votes based on the K-nearest neighbors, where neighbors are defined as the samples with the smallest Euclidean distance $||X_i - X_0||^2$, where X_i is the training data and X_0 is the test data point that has to be classified.

k-NN was self implemented in this project. A binary classifier was used to classify the samples into cancer versus non cancer samples and between Glioblastoma versus Oligodendroglioma and Astrocytoma, i.e., low grade versus high grade tumors. Multi-class classification was performed to classify the three subtypes of gliomas.

The best K value must be chosen such that it accurately captures and classifies the data. If K is too small it often becomes too sensitive to noise and if it is too high, all new samples will get classified into the same class thus there will be too much smoothing.

Best K for each experiment was chosen by calculating the accuracy for K in range $1 - 10$ for every experiment and selecting the value of K with the highest accuracy in each case. 5-fold cross-validation

was then performed using the best K to evaluate the performance of the model in terms of Accuracy, Precision, Recall and F1 scores.

2.4 Logistic Regression

Logistic regression is one of the most widely used classification methods in biology and for cancer classification. The goal of logistic regression is to classify a given input into binary classes, example a person has cancer or does not have cancer. It uses a sigmoid funtion to map an input to a class by assigning a weight to each input feature that represents its importance in classification. A stochastic gradient descent algorithm was used to minimize the loss function for random initializations for each training sample. The hyperparameter alpha was set to 1.25 for $10,000$ iterations.

$$h(x) = \frac{1}{1+e^{-}w^{T}x}$$

Logistic Regression was implemented without any packages for this project A binary classifier was used to distinguish between normal and glioma samples, and between low grade and high grade gliomas. As logistic regression is inherently a binary classifier, we used one-vs-all strategy for multi-class output. We created a one-hot vector vector where the correct class was assigned the value 1 and rest were assigned the value 0 for multiclass classification. The multinomial logistic classifier uses a generalization of the sigmoid, called the softmax function.

We also used L2 regularisation to prevent over fitting results in our model. L2 regularization corresponds to assuming that weights are distributed according to a Gaussian distribution with mean $\mu = 0$. The tuning parameter lambda (λ) was set to 0.01 to obtain the best results.

$$J(w) = \frac{1}{m}\sum_{i=1}^{m} Cost(h(x^{i}), yi) + \frac{\lambda}{2m}\sum_{j=1}^{n} w_{j}^{2}$$

2.5 Support Vector Machine

Support Vector Machines or SVM is a supervised machine learning classification algorithm. It aims to classify objects by finding a hyperplane that maximizes the margin between the two classes. After this training, the test examples are mapped into the space and predictions are made to classify them as belonging to one group or the other depending on which side of the hyperplane they belong to. Albeit, this case is for a linearly separable data. In order to implement SVM for non-linearly separable classes, the feature space is transformed to a higher dimensional space, where it becomes easier to fit the hyperplane as a decision boundary between classes. This is the basis for the 'kernel trick' in SVM. In this project, a built-in implementation of a kernalized SVM was employed under the 'scikit-learn' module. Radial basis function (RBF) was chosen as the function for the kernel trick as we expect the data to not be linearly separable. As SVM is not prone to outliers in classes, we expect SVM to perform best.

3 Results

All four classification methods were implemented after performing a dimensionality reduction on the number of genes using PCA. This was followed by a 5-fold cross validation, comparing the results of each classifier.

3.0.1 High performance metrics observed for Glioma vs Control classification:

To start the series of experiments, a simple binary classification of patients with Glioma and without Glioma was performed. A high separability of the two classes was observed with the help of a scatter plot of the first two dimensions from the results of PCA (Fig. **??**). The performance across LR, SVM and K-NN was consistent at 100% accuracy. This was followed by GNB that 94% accuracy. The lower performance for GNB is suspected to be because of the assumptions of conditional independence between genes, and normal distribution for gene expression that were made, which, in reality, may not be the case. The performance metrics are summarized in Table 1.

Table 1: Performance metrics of the 5-fold cross validation of Glioma vs Normal classification

Classifier	Accuracy	Precision	Recall	F1 Score
k-NN	100%	100%	100%	100%
GNB	94%	94%	94%	94%
LR	100%	100%	100%	100%
SVM	100%	100%	100%	100%

3.0.2 Decent performance on classification of the Glioma subtypes:

The classification of the subtypes attained the highest accuracy of 86% in the case of SVM, followed by LR (82%), K-NN (76%) and GNB (72%). SVM outperformed other classifiers in terms of precision (87%) and recall (85%) as well. The performance metrics is summarized in Table 2. A quick observation of the scatter plot of the first two dimensions from PCA (See Fig. **??**) reveal an overlap between two subtypes - Astrocytoma and Oligodendroglioma. Both these subtypes are of lower grade Glioma, often associated with benign cancers, and this was suspected to be a reason for the low separability between the two classes. This formed the motivation to classify between lower grade benign (Astrocytoma + Oligodendroglioma) and higher grade malignant (Glioblastoma).

Table 2: Performance metrics of the 5-fold cross validation of classifying the three subtypes of Glioma

Classifier	Accuracy	Precision	Recall	F1 Score
k-NN	76%	76%	75%	75%
GNB	72%	72%	72%	71%
LR	82%	81%	81%	81%
SVM	86%	87%	85%	85%

3.0.3 Better performance observed in classifying benign vs malignant:

By binning Astrocytes and Oligodendroglioma under one class (benign) and classifying it against Glioblastoma (malignant), we were able to better separate between the two classes in our PCA plots (See Fig. 1c), which resulted in better performance across all classifiers (See Table 3). This could indicate that differential genes between Astrocytes and Oligodendrocytes seem to be much lesser when compared to the differential genes between each of these two subtypes and Glioblastoma. In fact, a quick survey of literature reveals that the gene expression profiles seem to significantly overlap across these two subtypes[10].

Table 3: Performance metrics of the 5-fold cross validation of classifying benign vs malignant

Classifier	Accuracy	Precision	Recall	F1 Score
k-NN	94%	92%	94%	93%
GNB	83%	81%	81%	81%
LR	93%	92%	92%	92%
SVM	95%	95%	94%	94%

3.0.4 Sanity Checks - Separability of each Glioma subtype from Control samples:

As the final leg of our project, sanity checks were conducted to test the classifier for its ability to distinguish between the cancer sub-types as well as distinguish them from the non-cancerous samples. The PCA plots (Fig. 1d- 1e) produced much more separable clusters, as expected, and the overall performance of all metrics increased. This check was done for both cases - (a) Astrocytoma vs Oligodendroglioma vs Glioblastoma vs Control (See Table 4), and (b) Benign vs Malignant vs Control (See Table 5).

Table 4: Performance metrics of the 5-fold cross validation of classifying all three Glioma subtypes and Control samples

Classifier	Accuracy	Precision	Recall	F1 Score
k-NN	92%	89%	89%	89%
GNB	77%	77%	77%	76%
LR	84%	82%	82%	82%
SVM	89%	89%	88%	88%

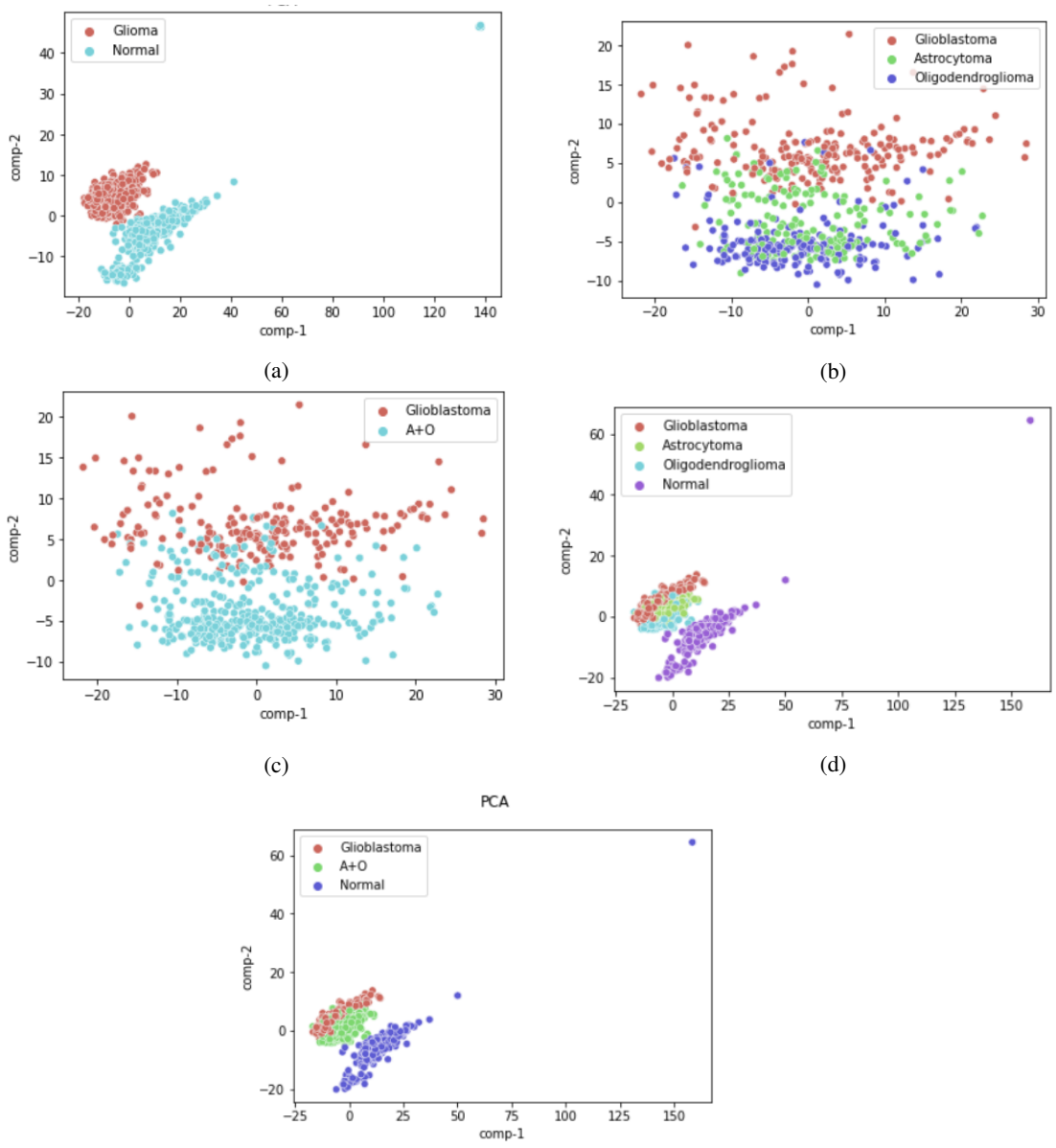

Figure 1: Scatter plot of genes with the first two dimensions from PCA analysis: (a) Glioma vs Control case, (b) Three Glioma subtypes case, (c) Benign vs Malignant case, (d) Sanity check 1: Three subtypes and Control case, (e) Sanity check 2: Benign vs Malignant vs Control case

Table 5: Performance metrics of the 5-fold cross validation of classifying benign vs malignant vs control

Classifier	Accuracy	Precision	Recall	F1 Score
k-NN	96%	94%	95%	95%
GNB	87%	86%	88%	87%
LR	94%	94%	94%	94%
SVM	97%	96%	96%	96%

4 Discussion & Conclusion

The overall best classifier for this experiment turned out to be SVM. Although, it seems like an expected result since we can observe a clear margin between classes in our PCA plots. k-NN and LR follow behind SVM, but by themselves they seem to exhibit decent classification performance metrics. An interesting observation to note would be that k-NN and LR seem to have comparable metrics in all analysis except the case where we classify between the three subtypes. This notable difference between the performance of k-NN and LR could be because k-NN depends on how close each of the data points in a class are to a test data point, while LR depends on whether the two classes as a whole are linearly separable or not. We observe a lot of overlap between the data points of Astrocytoma and Oligodendroglioma, despite which as a class, they appear somewhat linearly separable. We suspect this could be a reason for the lower classification metrics in k-NN but not in LR. Finally, GNB and its assumptions of conditional independence and normal distribution of gene expression leads to the classifier having subpar results compared to the other classifiers. As a rationale behind this, if we consider two features which are strongly correlated then in the case of GNB, both these features contribute equal influence over the probability distributions, thereby increasing both their importance in classification. By contrast, other classifiers seem much more suitable for correlated features as they either adjust assigning the weights to each feature (in case of SVM and LR) or just find the class of the nearest data points without any assumptions (in case of k-NN). By this logic, the SVM, LR and k-NN are definitely the better contenders.

As such, we evaluated the classification of Glioma subtypes using four commonly used classifiers. It would be interesting to observe how other complicated models like Feedforward Neural Networks perform with the data at hand. Another likely goal could be to expand on the feature selection or dimensionality reduction methods and conduct a cross-study with the "most important" genes from these techniques that help run the classification.

5 References

[1] Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., & Ellison, D. W. 2016. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary., *Acta neuropathologica, 131(6)*, pp. 803–820.

[2] Pollak, J., Rai, K.G., Funk, C.C., Arora, S., Lee, E., Zhu, J., Price, N.D., Paddison, P.J., Ramirez, J.M. and Rostomily, R.C., 2017. Ion channel expression patterns in glioblastoma stem cells with functional and therapeutic implications for malignancy. PLoS One, 12(3), pp.e0172884.

[3] Jin, X., Kim, L.J., Wu, Q., Wallace, L.C., Prager, B.C., Sanvoranart, T., Gimple, R.C., Wang, X., Mack, S.C., Miller, T.E. and Huang, P., 2017. Targeting glioma stem cells through combined BMI1 and EZH2 inhibition. Nature medicine, 23(11), pp.1352-1361.

[4] Ghosh, D., Ulasov, I.V., Chen, L., Harkins, L.E., Wallenborg, K., Hothi, P., Rostad, S., Hood, L. and Cobbs, C.S., 2016. TGF-responsive HMOX1 expression is associated with stemness and invasion in glioblastoma multiforme. Stem Cells, 34(9), pp.2276-2289.

[5] Louis, D.N., Holland, E.C. and Cairncross, J.G., 2001. Glioma classification: a molecular reappraisal. The American journal of pathology, 159(3), pp.779.

[6] Chakraborty, S., Mallick, B.K., Ghosh, D., Ghosh, M. and Dougherty, E., 2007. Gene expression-based glioma classification using hierarchical Bayesian vector machines. Sankhyā: The Indian Journal of Statistics, pp.514-547.

[7] Ranjith, G., Parvathy, R., Vikas, V., Chandrasekharan, K. and Nair, S., 2015. Machine learning methods for the classification of gliomas: Initial results using features extracted from MR spectroscopy. The neuroradiology journal, 28(2), pp.106-111.

[8] Way, G.P., Allaway, R.J., Bouley, S.J., Fadul, C.E., Sanchez, Y. and Greene, C.S., 2017. A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma. BMC genomics, 18(1), pp.1-11.

[9] Abusamra, H. (2013). A comparative study of feature selection and classification methods for gene expression data of glioma. Procedia Computer Science, 23, pp. 5-14. Chicago

[10] Hägerstrand, D., Smits, A., Eriksson, A., Sigurdardottir, S., Olofsson, T., Hartman, M., ... Ostman, A. (2008). Gene expression analyses of grade II gliomas and identification of rPTP/ as a candidate oligodendroglioma marker. Neuro-oncology, 10(1), pp 2-9.