Artificial Intelligence in Oncology

Available online at: www.aioncology.org

Original Article

An Overview of Mathematical Models for RNA Sequence-based Glioblastoma Subclassification

Eric Zander, Andrew Ardeleanu, Ryan Singleton, Barnabas Bede, Ph.D., Yilin Wu, Ph.D.*

DigiPen Institute of Technology, Department of Mathematics, Washington, 98052, U.S.A.

ARTICLE INFO

Article history:
Received 14 April 2021
Revised 21 May 2021
Accepted 22 May 2021
Published 01 June 2021

Keywords:
Glioblastoma multiforme,
random forest,
machine learning,
mathematical model.

ABSTRACT

Molecular marker-based glioblastoma (GBM) subclassification is emerging as a key factor in personalized GBM treatment planning. Multiple genetic alterations, including methylation status and mutations, have been proposed in GBM subclassification. RNA-Sequence (RNA-Seq)-based molecular profiling of GBM is widely implemented and readily quantifiable. Machine learning (ML) algorithms have been reported as an applicable method that can consistently subgroup GBM. In this study, we systematically studied the applicability of the commonly used ML algorithms based on The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) dataset and cross-validated in the Chinese Glioma Genome Atlas (CGGA) dataset. ML algorithms studied include Binomial and multinomial Logistic Regression, Linear discriminant analysis, Decision trees, K-Nearest Neighbors, Gaussian Naive Bayes, Support Vector Machine, Random Forest, Gradient Boosting, Multi-Layer Perceptron, and Voting Ensemble. We found Support Vector Machine, Multi-Layer Perceptron, and Voting Ensemble best equipped to assign GBM to correct molecular subgroups of GBM without histological studies.

Introduction

A study carried out by Verhaak *et al.* (2010) provided evidence of distinct and clinically relevant 4 subtypes of glioblastoma multiforme (GBM) distinguishable by genomic abnormalities [1]. These include a Mesenchymal subtype defined by NF1 mutation, a Classical subtype with notable *EGFR* abnormalities, a Proneural subtype typified by distinct *PDGFRA* and *IDH1* events, and a Neural subtype defined by abnormal expressions of neuron markers such as NEFL [1, 2]. Each of these also exhibited unique alterations and mutations in numerous genetic markers [1, 2]. More recent studies proposed that the Neural subtype arose from contamination of the nonmalignant samples, and therefore, GBM is now grouped into three subtypes, i.e., Mesenchymal, Classical and Proneural[3].

Emerging studies indicated the clinical significance of these subclassifications in predicting treatment effectiveness [2, 4-6]. Cases describable by each of the 3 subtypes respond uniquely to therapies and may guide personalized treatment [6]. The Classical and Proneural subtypes correspond with better therapeutic response than the Mesenchymal subtype, which often exhibits more aggressive features than other subtypes [1, 7]. Meanwhile, the Classical and Mesenchymal subtypes may benefit from more intensive treatment regimens (concurrent chemo-/radio-therapy or greater than 4 cycles of chemotherapy) [1]. Given this, reliably identification of GBM subtypes is of exceptional importance for personalized GBM treatment. Complicated genetic markers such as mutations and methylation identifying the individual subtype with high accuracy have been proposed [8]. However, identifying patients likely to benefit from intensive treatment for improved overall survival (OS) rates constitutes a more straightforward and clinically significant binary

* Corresponding author: Yilin Wu, Ph.D.
E-mail address: y.wu@digipen.edu

classification problem.

Leveraging genetic expression levels to predict GBM subtype or likely response to intensive treatment can be accomplished by employing one of many available machine learning (ML) classification algorithms. Potentially applicable models include probabilistic models, tree-based classifiers, connectionist models, ensemble models, and more. Existing studies highlight the potential value of employing ML models with RNA-seq data to solve cancer classification tasks, and suggest the relative effectiveness of support vector machines (SVMs) and ensemble models such as random forests (RF) in some contexts [9, 10]. However, further research also illustrates the value of gradient boosting and connectionist models in solving problems with similar quantitative data [5, 11]. GBM subtype classification with gradient boosting classifiers and radiomics data has also yielded encouraging results, suggesting both the potential applicability of boosting models and the importance of bolstering systematic subtype classification efforts [12]. To identify the most appropriate models for GBM subtype classification, we evaluated the performance of these approaches as well as numerous other traditionally employed model types for comparison.

This study aims to make advances toward the goal of reliably labelling subtypes based on continuous genetic data while comparing the applicability of various options in this context. To achieve this goal, numerous popular classification models are trained and evaluated with metrics that account for class imbalances. Cross-validation is then performed across datasets to verify the consistency of each model. Consequently, this study constitutes an overview and experimental comparison of mathematical algorithms for GBM subclassification based on RNA-sequence (RNA-Seq) of key genetic markers.

Methods and Materials

Public datasets

To train and experimentally compare models prior to cross-validation, The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) dataset has been used. For cross-validation, models have been tested on GBM data from the Chinese Glioma Genome Atlas (CGGA) dataset [13]. Both are available through the GlioVis platform (http://gliovis.bioinfo.cnio.es/, last accessed April 10th, 2021). Basic patient information from TCGA-GBM and CGGA datasets that are included in the study in presented in **Table 1,2**. While each set includes numerous gene expression levels features and labeled subtypes, the labels in the CGGA set only describe Mesenchymal, Classical, and Proneural subtypes. The TCGA set has two separate label features describing all four subtypes and only these three subtypes.

Table 1: Basic patient demographics from TCGA dataset

Subgroups	Cases	Gender (M : F)	IDH Mutation	Age (Avg.)
Mesenchymal	50 (32.3%)	28 : 22 (56.0 : 44.0%)	0 (0.0%)	61.7
Classical	40 (25.8%)	26 : 14 (65.0 : 35.0%)	0 (0.0%)	60.9
Proneural	38 (24.5%)	27 : 11 (71.1 : 28.9%)	8 (21.1%)	61.9
Neural	27 (17.4%)	20 : 7 (74.1 : 25.9%)	0 (0.0%)	57.7

Table 2: Basic patient demographics from CGGA dataset

Subgroups	Cases	Gender (M : F)	IDH Mutation	Age (Avg.)
Mesenchymal	51 (32.9%)	31 : 20 (60.8 : 39.2%)	0 (0.0%)	61.6
Classical	59 (38.1%)	37 : 22 (62.7 : 37.3%)	0 (0.0%)	60.8
Proneural	45 (29.0%)	33 : 12 (73.3 : 26.7%)	8 (17.8%)	58.7

Table 3: Cross-validation with CGGA dataset

Fold	LR	LDA	DT	KNN	GNB	SVC LIN	SVC RBF	RF	GBOOST	MLPNN	VOTE
1	0.77	0.64	0.63	0.86	0.76	0.8	0.79	0.77	0.74	0.77	0.8
2	0.81	0.64	0.8	0.73	0.84	0.78	0.76	0.84	0.84	0.68	0.78
3	0.9	0.49	0.71	0.81	0.87	0.81	0.77	0.87	0.74	0.87	0.9
4	0.77	0.71	0.41	0.81	0.64	0.74	0.7	0.71	0.65	0.73	0.74
5	0.55	0.48	0.44	0.49	0.44	0.5	0.51	0.47	0.54	0.53	0.51

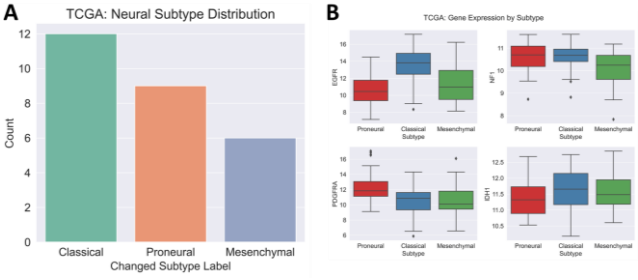

Figure 1: TCGA-GBM Subtype Descriptions. A. A histogram of subtype label counts in absence of the Neural classification show that of the 27 cases previously labelled Neural, 12 are Classical, 9 are Proneural, and 6 are Mesenchymal. **B**. Expression levels of genes highlighted by Verhaak *et al*. as distinct among subtypes remain so in absence of the Neural subtype. The results presented here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Classification was performed on the feature lacking the Neural subclassification.

Patients included in the study

The TCGA dataset includes cases of patients with the following subtype, gender, *IDH* mutation, and age distributions. The cases labeled Neural with all subclassifications present are split as in **Figure 1A, B**. The patterns of abnormalities in *EGFR*, *NF1*, *PDGFRA*, and *IDH1* expression levels described by Verhaak *et al*. hold true in the absence of the Neural subtype. Classical cases see high *EGFR* levels, Mesenchymal cases see low *NF1* levels, and Proneural cases see unique *PDGFRA*/*IDH1* levels. The CGGA

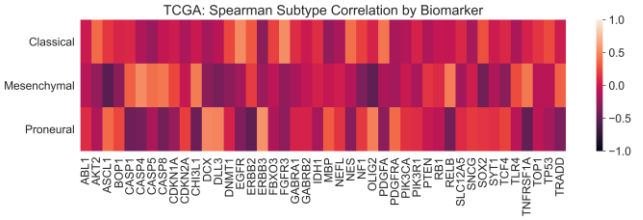

Figure 2: Heatmap of Spearman's Rank Correlation Coefficients for 44 Biomarkers by GBM Subtype. The monotonic relationship between subtype and expression levels of genes described as distinct by Verhaak *et al*. illustrate how subtypes may be differentiated using genetic data. Among these, markers such as *ASCL1*, *CHI3L1*, *EGFR*, *ERBB2*, *OLIG2*, and *RELB* stand out as potentially valuable for classification.

dataset used for cross-validation includes the same expression level features alongside the following distribution (**Table 3**).

Biomarkers selection for GBM subclassification

Verhaak *et al*. highlight differences in gene expression of *EGFR*, *NF1*, *PDGFRA*, and *IDH1*, but also describe mutations in *TP53*, *RB1*, *PIK3R1*, *ERBB2*, and numerous other potential biomarkers [1]. To improve model performance, a total of up to 44 biomarkers are employed in training and predicting subtypes. A heatmap of Spearman rank's correlation coefficients for each biomarker by subtype (**Figure 2**).

Algorithms for GBM subclassification

The models considered and compared in this study are available from the scikit-learn package [14, 15]. While hyperparameter selection varied for each model type, 5-fold cross-validation within the TCGA dataset has been performed to select values when not specified. Included here are brief summaries of each model type. Data is scaled or standardized as required for models requiring comparable scales or that predictors may be described by a standard distribution. Model performances are then evaluated by calculating macro-averaged F-scores to account for subtype representation imbalances exhibited by both the TCGA and CGGA datasets. Average accuracy, precision, and recall are also calculated and compared for a more holistic understanding of model performance.

1. *Binomial and multinomial logistic regression (LR).* LRs are statistical classification models that leverage a logistic representation of data to predict the probability of cases belonging to various classes [16]. Performing logistic regression involves optimizing a cost function and requires regularization, as excluding unlikely values can combat overfitting that would otherwise deteriorate performance in the context of high-dimensional data. The sklearn model used employs L2 regularization:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

2. *Linear discriminant analysis (LDA).* LDA also employs probabilistic models [17]. New data is classified by generating linear boundaries between inputs and determining the probability of a novel vector belonging in each defined region. In the case of sci-kit learn's implementation, probabilities are calculated with the assumption that prediction data conforms to a Gaussian distribution:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

3. *Decision trees (DTs).* DTs learn by splitting data into subsets based on the features that result in the highest class purity following each split [18]. Even after selecting preferable stopping criteria and pruning methods designed to minimize overfitting, single decision trees are often referred to as weak learners due to poor performance. However, they are fast, easily interpretable, and can be leveraged effectively in ensemble models; both the random forest and gradient boosting models tested in this study involve classification by majority vote with numerous distinct decision trees. An example built with a max depth of 3 using Gini impurity as a metric splits on *FGFR3*, *ASCL1*, *SNCG*, *CASP5*, *ERBB2*, and *ERBB3*; dividing training data on these features results in leaves with the highest subclass purities (**Figure 3**).

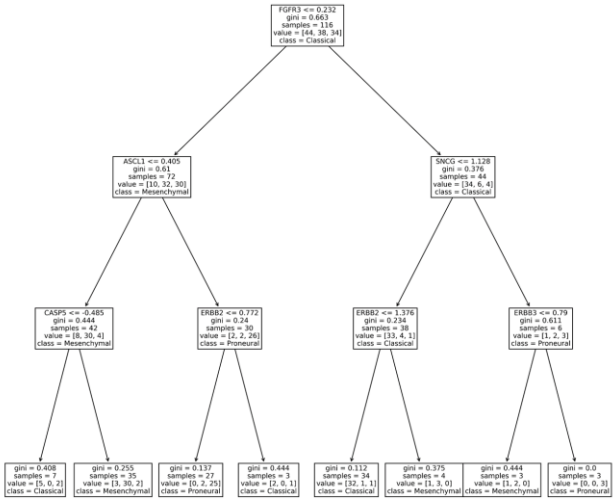

Figure 3: Example Decision Tree Classifier Derived with Gini Impurity. Creating a decision tree with a max depth of 3 results in splits along *FGFR3*, *ASCL1*, *SNCG*, *CASP5*, *ERBB2*, and *ERBB3* expression levels. These result in the highest subtype purities after splitting. While using single decision trees such as this can lead to suboptimal performances, employing multiple different tree classifiers in ensemble models can prove quite effective for classification tasks such as this.

4. *K-Nearest Neighbors (KNN).* KNN is a simple model that calculates distances between a given vector and all training points [19]. The majority class of a given number of closest points is returned as the prediction. While efficient to train, KNN can struggle with high dimensional data. K-d trees are employed in this context.

5. *Gaussian Naive Bayes (GNB).* GNB classifiers are probabilistic models. Although they assume the conditional independence of predictors, they can sometimes perform surprisingly well even if this assumption is false [20]. To make predictions, the likelihoods of a case belonging to each subclass are calculated and compared. These likelihoods are products of the prior probability and the likelihoods for each individual predictor. Gaussian

Naive Bayes determines these individual likelihoods for continuous features by assuming each feature belongs to a standard distribution:

$$P(x|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)\right)$$

6. *Support Vector Machines (SVMs).* SVMs are kernel-based models that create hyper-planes to separate data based on subclassification. While there are numerous usable kernel functions, the popular linear (SVCLIN) and radial basis (SVCRBF) kernels are tested here.

7. *Random forests (RFs).* RFs are bootstrap aggregating ensemble models that create numerous decision trees. Each tree is trained using random subsets of the training data sampled with replacement, ideally resulting in many weak learners with unique strengths and weaknesses. Predictions are then made using each tree, and the most common result serves as the final prediction.

8. *Gradient Boosting (GBOOST).* Boosting models are ensemble models consisting of many weak learners. Decision trees are also used in this case. Each created tree is subsequently built to optimize a loss function. This ideally results in many unique and increasingly effective trees. However, this additional optimization also has the capacity to render GBOOST models subject to overfitting.

9. *Multi-Layer Perceptron (MLP).* MLP was also studied to represent connectionist approaches to classification. Neural networks are flexible and potent, but with the variety of potentially performant architectures, activation functions, and numerous other options for further optimization comes additional complexity and a lack of interpretability. In this case, MLPs with a single hidden layer of 10 neurons have provided a relatively superior balance of performance and simplicity in the context of cross validation within the TCGA dataset. Consequently, this architecture underlies the results for classification described in this comparison. Stochastic gradient descent is used to train the network.

10. *Voting Ensemble (VOTE).* VOTE models can be created using other classification models. Ideally, taking the majority vote of multiple unique performant models can result in greater accuracy should the used models compensate for each others' weaknesses. In this case, a voting model employing the most performant instance of each of the other model types is tested.

Results

Three Subtypes Training and Testing with TCGA Dataset

The accuracies of five-fold cross validation for the most performant of each model type are as presented in **Table 3**. The average classification F-score across models is 0.702. The best average F-score for a single model, attributable to logistic regression (LR), is 0.759. Notably, LR models trained on data with many dimensions can fall prey to overfitting; cross validation results will reveal that the voting model, with an F-score of 0.747, proves more consistent. Accuracies by subtype are presented in **Figure 4A, B**. While models can more reliably identify a Classical subtype, they universally have difficulties correctly identifying Proneural subtypes. The misclassified Proneural subtypes are predicted to be Classical and Mesenchymal with similar frequencies (**Figure 5C**).

Three Subtypes Cross Validation with CGGA Dataset

Accuracies of predictions made for the CGGA dataset using model trained on the TCGA dataset are presented in **Table 4**. Models such as LR had greater drops in accuracy than SVMs, the MLP, and the ensemble models, but all types saw a drop in accuracy across the board. The average accuracy was 59.86%, and the best accuracy, attributable to the SVM employing a linear kernel function, was 63.39%. Interestingly, we found accuracy by subtype reveals a similar pattern to the one observable in the context of the TCGA dataset (**Figure 5 A, B**). This data also indicate Classical subtype in

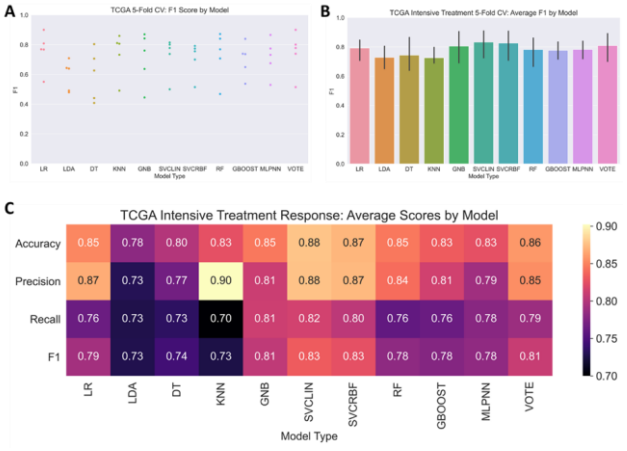

Figure 4: TCGA-GBM 5-Fold Cross Validation Results by Classification Model. A. 5-Fold cross validation F-scores reveal the in-set applicability and consistency of each model type. **B.** Visualizing average cross validation scores reveals the value of LR, KNN, SVM, MLPNN, and ensemble models within the dataset. **C.** Model precision scores tend to be higher than recall scores.

Table 4: Prediction of CGGA dataset GBM subclasses with models trained with TCGA dataset

LR	LDA	DT	KNN	GNB	SVC LIN	SVC RBF	RF	GBOOST	MLPNN	VOTE
0.58	0.58	0.47	0.59	0.62	0.63	0.6	0.59	0.58	0.61	0.6

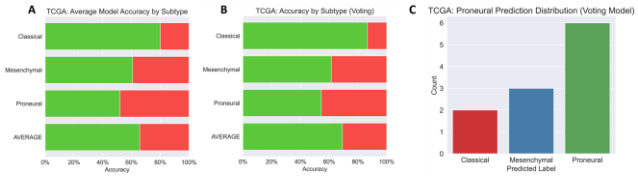

Figure 5: TCGA-GBM Model Accuracies by Subtype. A. Examining the average accuracy by subtype of all model types illustrates high success in correctly identifying cases labelled Classical, but much lower consistency in correctly labelling Proneural cases. **B.** Observing accuracies by subtype for just the ensemble voting model reflects the higher overall accuracy of the model as well as the pattern of difficulty with identifying cases labelled as the Proneural subtype. **C.** Though the number of proneural cases in the testing cohort is low, the nearly even distribution of incorrect predictions should be noted.

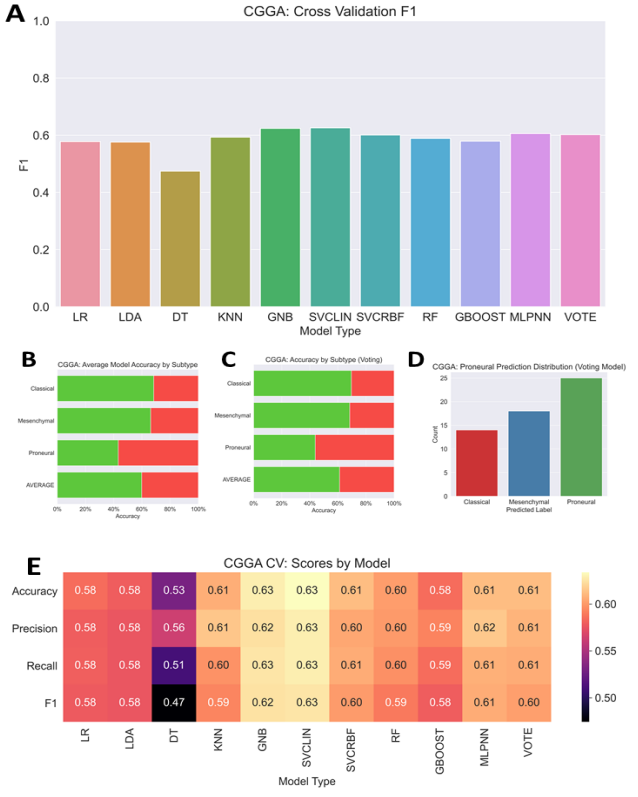

Figure 6: CGGA GBM Cross Validation Results. A. F-scores drop when cross validating with the CGGA dataset. This is most severe for LR, which can suffer from overfitting with high dimensional data. **B.** The average cross validation accuracy by subtype across models shows a similar pattern seen within the TCGA dataset, with low Proneural accuracy. **C.** Accuracy by subtype for the voting ensemble model further reflects issues with Proneural cases. **D**. As with the TCGA set, model recall scores tend to be lower across models. **E**. Mislabelled Proneural cases tend to be classified as Mesenchymal more often by the voting ensemble model.

Table 5: F-score of models trained with TCGA dataset for intensive treatment

Fold	LR	LDA	DT	KNN	GNB	SVC LIN	SVC RBF	RF	GBOOST	MLPNN	VOTE
1	0.82	0.72	0.8	0.69	0.83	0.82	0.82	0.78	0.78	0.8	0.82
2	0.87	0.81	0.72	0.69	0.81	0.88	0.88	0.87	0.81	0.81	0.88
3	0.82	0.64	0.96	0.69	0.96	0.92	0.92	0.87	0.88	0.77	0.92
4	0.83	0.84	0.58	0.87	0.82	0.92	0.92	0.83	0.72	0.88	0.83
5	0.63	0.64	0.66	0.69	0.6	0.63	0.6	0.56	0.69	0.66	0.6

Table 6: F-score of models trained with CGGA dataset for intensive treatment

LR	LDA	DT	KNN	GNB	SVC LIN	SVC RBF	RF	GBOOST	MLPNN	VOTE
0.63	0.66	0.59	0.54	0.66	0.65	0.69	0.63	0.64	0.68	0.65

general is more easily identified, while Proneural subtypes are more easily misidentified (**Figure 5 A, B**).

Intensive Treatment Response Training and Testing

Predicting the likely response to intensive treatment based on subtype group rather than individual subtype leads to an unsurprising increase in accuracy in the TCGA-GBM dataset. The average F-score across models is now 0.782, and the SVM with a linear kernel predicted subtype category with an F-score of 0.832 (**Table 5**, **Figure 6**). Intensive treatment response was cross validated in the CGGA dataset (**Table 6**, **Figure 7**). Although drops in accuracy are also observable, the scores are also higher. On average, the models predicted prospective treatment response with an F-score of 0.631. The support vector machine using the radial basis function kernel performed best with an accuracy of 0.686.

Conclusion

Our study showed that F-scores around 0.750 for the training set and just over 0.600 for the cross-validation far outclasses guessing, but still leaves much to be desired. We suspected that overlap between biomarkers as well as the limited number of cases with the relevant data within these sets constitutes partial explanations for why these scores fall short. Another possible explanation involves potential systematic differences in the demographics and determination of expression levels between the two datasets. This may also contribute to the gap in performance. Simplifying the problem by classifying prospective response to intensive treatment does improve these scores to over 0.800 for the training set and over 0.675 for the testing set, but at the cost of identifying the individual subtype. To better understand these metrics and how they may be improved, a closer look can be taken at performance by model type.

Outside of single decision trees and linear discriminant analysis, no models performed particularly well or poorly in comparison to the rest. This constitutes the fundamental takeaway regarding comparisons of models; the similar performance across the board indicates that the inability to reach higher accuracies may primarily be attributable to factors beyond model selection. Additionally, the lack of improvement seen in ensemble models and the voting model in particular can be explained by a lack of unique weaknesses or strengths among the different model types in this context.

However, the final scores, while similar, have some differences worth consideration.

Following cross-validation with the CGGA dataset, it becomes clear that SVMs, MLPs, and voting ensemble models ultimately performed best. The relatively poor performance of some models may be explained by their weaknesses. Though some regularization has been performed prior to cross-validation between datasets, LR models can suffer from overfitting. GNB models must assume conditional independence between the biomarkers used for prediction. While the tenuous nature of this assumption in this context may slightly lower performances, GNB performed decently enough in cross-validation to warrant recognition. However, DTs and KNN models, while easily interpreted, are ostensibly unable to outperform other more elaborate models for this particular problem. Such shortcomings afford the superior performances of SVMs, connectionist models, and ensemble models both in literature concerning classification of RNA-seq data and here. Determining the reasons underlying the gaps in performance between these models proves more difficult given the largely similar outcomes.

The relatively high performance of SVMs may be attributable to the ability to perform well with high dimensional data. This may partially explain the slightly superior results in cross-validation with the CGGA dataset when compared to tree-based ensemble models. While both RF and GBOOST models outperformed the weakest models, they achieved marginally lower scores in cross-validation. Particularly in the case of GBOOST, the higher relative scores prior to cross-validation may indicate some degree of overfitting. While this can also present a problem for MLP and VOTE models, their performances and the latter's ability to balance the weaknesses of its estimators suggest distinct value in evaluating RNA-seq data. Fundamentally though, the similarity of scores attributable to all of the more performant models indicate that the key to real improvement in subtype classification will require work beyond theoretically informed model selection.

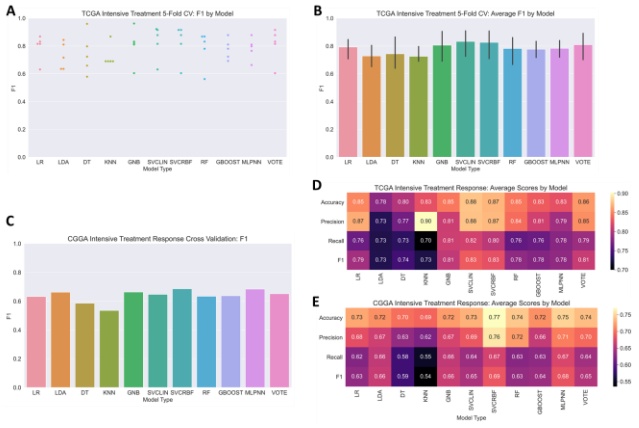

Figure 7: TCGA-GBM Model F-Scores and CGGA Cross Validation Results for Predicting Response to Intensive Treatment. A. Examining F-Scores for the TCGA set illustrates higher and more consistent results relative to classifying individual subtype. **B.** Visualizing average F-score by model clarifies which models are more performant. SVMs and ensemble models stand out in this regard. **C.** Cross validating with the CGGA dataset shows that SVMs with the radial basis kernel function perform best. **D.** Scores for each model reveal lower recall scores and highlight the relatively high performance of SVMs. **E.** CGGA scores show the value of the SVM with the radial basis function. Additionally, the MLPNN also performed relatively well.

Other avenues to improving model performances in the future could include further examination of individual subtype accuracies. Though Proneural cases were often misidentified as both the Classical and Mesenchymal subtypes, the slightly higher rate of misclassifications as Mesenchymal suggests that emphasizing the differences between Proneural and Mesenchymal subtypes may prove valuable in future endeavors; further preprocessing may be tailored to limit Proneural misclassification specifically. However, increasing data quantity constitutes the most promising path to substantial performance improvements. Particularly when employing demonstrably promising models like MLP, additional cases may be necessary for reaching the desired consistency.

Although tuning models and experimenting with further preprocessing could increase overall performance, the consistency of prediction accuracy across models indicates that the keys to a sizable improvement in performance lie elsewhere. That said, the evidence indicates that SVMs, MLPs, and ensemble models are best equipped to handle both classification problems with continuous genetic data in this context.

References

1. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 17: 98-110. doi: 10.1016/j.ccr.2009.12.020.
2. Dejaegher J, Solie L, Hunin Z, Sciot R, Capper D, Siewert C, Van Cauter S, Wilms G, van Loon J, Ectors N, Fieuws S, et al. (2021). DNA methylation based glioblastoma subclassification is related to tumoral T-cell infiltration and patient survival. Neuro Oncol. 23: 240-250. doi: 10.1093/neuonc/noaa247.
3. Sidaway P. (2017). CNS cancer: Glioblastoma subtypes revisited. Nat Rev Clin Oncol. 14: 587. doi: 10.1038/nrclinonc.2017.122.
4. Conroy S, Kruyt FA, Joseph JV, Balasubramaniyan V, Bhat KP, Wagemakers M, Enting RH, Walenkamp AM, den Dunnen WF. (2014). Subclassification of newly diagnosed glioblastomas through an immunohistochemical approach. PLoS One. 9: e115687. doi: 10.1371/journal.pone.0115687.
5. Khalifa NEM, Taha MHN, Ezzat Ali D, Slowik A, Hassanien AE. (2020). Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach. in IEEE Access. 8: 22874-22883. doi: 10.1109/ACCESS.2020.2970210.
6. Prados MD, Byron SA, Tran NL, Phillips JJ, Molinaro AM, Ligon KL, Wen PY, Kuhn JG, Mellinghoff IK, de Groot JF, Colman H, et al. (2015). Toward precision medicine in glioblastoma: the promise and the challenges. Neuro Oncol. 17: 1051-1063. doi: 10.1093/neuonc/nov031.
7. Azam Z, To ST, Tannous BA. (2020). Mesenchymal Transformation: The Rosetta Stone of Glioblastoma Pathogenesis and Therapy Resistance. Adv Sci (Weinh). 7: 2002015. doi: 10.1002/advs.202002015.

8. Ma H, Zhao C, Zhao Z, Hu L, Ye F, Wang H, Fang Z, Wu Y, Chen X. (2020). Specific glioblastoma multiforme prognostic-subtype distinctions based on DNA methylation patterns. Cancer Gene Ther. 27: 702-714. doi: 10.1038/s41417-019-0142-6.
9. Johnson NT, Dhroso A, Hughes KJ, Korkin D. (2018). Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? RNA. 24: 1119-1132. doi: 10.1261/rna.062802.117.
10. Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Zararsiz GE, Duru IP, Ozturk A. (2017). A comprehensive simulation study on classification of RNA-Seq data. PLoS One. 12: e0182507. doi: 10.1371/journal.pone.0182507.
11. Ma B, Meng F, Yan G, Yan H, Chai B, Song F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med. 121: 103761. doi: 10.1016/j.compbiomed.2020.103761.
12. Le NQK, Hung TNK, Do DT, Lam LHT, Dang LH, Huynh TT. (2021). Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from MRI. Comput Biol Med. 132: 104320. doi: 10.1016/j.compbiomed.2021.104320.
13. Zhao Z, Zhang KN, Wang Q, Li G, Zeng F, Zhang Y, Wu F, Chai R, Wang Z, Zhang C, Zhang W, et al. (2021). Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Gliomas. Genomics Proteomics Bioinformatics. doi: 10.1016/j.gpb.2020.10.005.
14. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G. (2014). Machine learning for neuroimaging with scikit-learn. Front Neuroinform. 8: 14. doi: 10.3389/fninf.2014.00014.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach Learn Res. 12: 2825-2830.
16. Singh V, Dwivedi SN, Deo SVS. (2020). Ordinal logistic regression model describing factors associated with extent of nodal involvement in oral cancer patients and its prospective validation. BMC Med Res Methodol. 20: 95. doi: 10.1186/s12874-020-00985-1.
17. Hu W, Shen W, Zhou H, Kong D. (2020). Matrix Linear Discriminant Analysis. Technometrics. 62: 196-205. doi:
18. Che D, Liu Q, Rasheed K, Tao X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. Adv Exp Med Biol. 696: 191-199. doi: 10.1007/978-1-4419-7046-6_19.
19. Zhang Z. (2016). Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 4: 218. doi: 10.21037/atm.2016.03.37.
20. Huang HJ, Hsu CN. (2002). Bayesian classification for data from the same unknown class. IEEE Trans Syst Man Cybern B Cybern. 32: 137-145. doi: 10.1109/3477.990870.