Joint Modeling of RNAseq and Radiomics Data for Glioma Molecular Characterization and Prediction

Zeina A. Shboul  
*Old Dominion University, zshbo001@odu.edu*

Norou Diawara  
*Old Dominion University, ndiawara@odu.edu*

Arastoo Vossough  
James Y. Chen  
Khan M. Iftekharuddin  
*Old Dominion University, kiftekha@odu.edu*

Follow this and additional works at: [https://digitalcommons.odu.edu/ece_fac_pubs](https://digitalcommons.odu.edu/ece_fac_pubs)

Part of the Biomedical Commons, Genetics Commons, Mathematics Commons, and the Radiology Commons

Original Publication Citation  
Shboul, Z. A., Diawara, N., Vossough, A., Chen, J. Y., & Iftekharuddin, K. M. (2021). Joint modeling of RNAseq and radiomics data for glioma molecular characterization and prediction. *Frontiers in Medicine*, 8, 1-13, Article 705071. [https://doi.org/10.3389/fmed.2021.705071](https://doi.org/10.3389/fmed.2021.705071)

This Article is brought to you for free and open access by the Electrical & Computer Engineering at ODU Digital Commons. It has been accepted for inclusion in Electrical & Computer Engineering Faculty Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
Joint Modeling of RNAseq and Radiomics Data for Glioma Molecular Characterization and Prediction

Zeina A. Shboul¹, Norou Diawara², Arastoo Vossough³, James Y. Chen⁴ and Khan M. Iftekharuddin*¹

¹ Vision Lab, Department of Electrical & Computer Engineering, Old Dominion University, Norfolk, VA, United States,
² Department of Mathematics & Statistics, Old Dominion University, Norfolk, VA, United States,
³ Department of Radiology, Children’s Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, United States,
⁴ University of California, San Diego Health System, San Diego, CA, United States

RNA sequencing (RNAseq) is a recent technology that profiles gene expression by measuring the relative frequency of the RNAseq reads. RNAseq read counts data is increasingly used in oncologic care and while radiology features (radiomics) have also been gaining utility in radiology practice such as disease diagnosis, monitoring, and treatment planning. However, contemporary literature lacks appropriate RNA-radiomics (henceforth, radiogenomics) joint modeling where RNAseq distribution is adaptive and also preserves the nature of RNAseq read counts data for glioma grading and prediction. The Negative Binomial (NB) distribution may be useful to model RNAseq read counts data that addresses potential shortcomings. In this study, we propose a novel radiogenomics-NB model for glioma grading and prediction. Our radiogenomics-NB model is developed based on differentially expressed RNAseq and selected radiomics/volumetric features which characterize tumor volume and sub-regions. The NB distribution is fitted to RNAseq counts data, and a log-linear regression model is assumed to link between the estimated NB mean and radiomics. Three radiogenomics-NB molecular mutation models (e.g., IDH mutation, 1p/19q codeletion, and ATRX mutation) are investigated. Additionally, we explore gender-specific effects on the radiogenomics-NB models. Finally, we compare the performance of the proposed three mutation prediction radiogenomics-NB models with different well-known methods in the literature: Negative Binomial Linear Discriminant Analysis (NBLDA), differentially expressed RNAseq with Random Forest (RF-genomics), radiomics and differentially expressed RNAseq with Random Forest (RF-radiogenomics), and Voom-based count transformation combined with the nearest shrinkage classifier (VoomNSC). Our analysis shows that the proposed radiogenomics-NB model significantly outperforms (ANOVA test, \( p < 0.05 \)) for prediction of IDH and ATRX mutations and offers similar performance for prediction of 1p/19q codeletion, when compared to the competing models in the literature, respectively.

Keywords: RNA sequencing, radiomics, radiogenomics, negative binomial, molecular mutation
INTRODUCTION

Radiomics is increasingly being applied to radiology practice in disease diagnosis, grading, monitoring, and treatment planning (1, 2). Radiomics is extracted from various radiological images of a targeted area of the disease. Fusing the important radiomics and genomics information in the proper computational machine learning (ML) model may help to achieve a more comprehensive disease diagnosis, prognosis, and treatment planning scheme (3–5). Different studies have evaluated the association between glioma molecular subtypes and radiomics (e.g., tumor shape and size) (6–8), or between different form of genomics (e.g., RNA sequencing (RNAseq) gene expression, protein expression, copy number, molecular mutations, or DNA methylation) and glioma subtypes (9–11).

Conventional ML models do not adequately model the count-based nature of the RNA-sequence data as these models are usually designed to work with data that has a normal distribution. In order to alleviate the lack of appropriate ML models, researchers propose to transform the RNAseq read-count data to approximate a normal distribution. The transformation to normal distribution allows the use of existing methods such as the nearest shrinkage method (12, 13) or Random Forest for classification. However, such transformation removes the count-based nature of the RNAseq read counts data, and hence, lacks the ability to fully preserve the strong mean-variance relationship that is otherwise useful for glioma classification and prediction (14, 15). In order to appropriately model RNAseq read-count data, Negative Binomial (NB) and Poisson distributions are commonly used (16). The Poisson distribution is a single parameter distribution with its mean equals to its variance, which makes it rather restrictive. On the other hand, NB is similar to a Poisson distribution with an additional parameter called “dispersion” that allows the NB distribution to modify its variance without affecting the mean.

RNAseq uses high-throughput or next-generation sequencing technology (NGS) and has emerged as a novel alternative to microarray-based techniques for quantifying gene expression. The microarray technique is known to suffer from background noise. Gene expression level is measured as the relative frequency of the RNAseq reads that are mapped to one gene (17). RNAseq is a very sensitive technique that provides high resolution and a thorough understanding of the transcriptome and has revealed many novel gene structures.

RNAseq distribution requires an appropriate model that adapts and preserves the nature of RNAseq read counts data, and such classification models that preserve the nature of RNAseq are lacking in the traditional ML literature. The NB distribution is an appropriate choice to model such discrete reads counts data (16). Even though traditional ML tools that are developed based on NB are lacking, the choice of using NB distribution in differential gene expression and RNAseq analysis has been adapted by different studies in the literature such as in EdgeR (18–20), DESeq (21), and NBSeq (22).

An example of a count-based classifier that fits a NB distribution is the Negative Binomial Linear Discriminant Analysis (NBLDA). NBLDA is a well-known classifier that is developed by fitting NB to RNAseq and the mean and dispersion parameter are estimated from the RNAseq data (23). A different type of classifier, known as VoomNSC, is developed based on the transformed count data. VoomNSC is a combination of Voom (an acronym for mean-variance modeling at the observational level) transformation (12) and the nearest shrunken centroids classifier (NSC) (24).

Consequently, the aim of this work is to implement a joint radiogenomics-NB model that predicts and classifies glioma molecular mutations following the 2016 World Health Organization’s (WHO) updated guidelines for classification of tumors of the Central Nervous System (CNS) (including high grade and diffuse low-grade gliomas) (25). This work is critical especially when the RNAseq of some cases are unknown and a careful assessment is needed to avoid mischaracterization of lower grade gliomas. In this work, we utilize both volumetric features (radiomics) and RNAseq to implement and learn a radiogenomics-NB model. Then, the trained radiogenomics model is used to predict and classify the unknown RNAseq data.

Figure 1B, the estimated parameters of regression coefficients and the dispersion parameters along with...
the sample radiomics and its RNAseq read counts are utilized to predict the class label of a future test sample. A complete mathematical derivation of the radiogenomics-NB model is presented in the following subsection.

**Prediction Using Negative Binomial Regression Model**

To fuse radiomics with RNAseq read counts data in an NB model, the following parametrization is defined:

Let $C$ be the total number of classes, and $I_c \in (1, \ldots, n_c)$ be the indices of samples in class $c$ for $c = 1, \ldots, C$. The examples of different classes include:

- IDH mutated vs. wildtype IDH ($C = 2$),
- 1p/19q codeletion: codeletion vs. non-codeletion ($C = 2$),
- Mutated ATRX vs. wildtype ($C = 2$).

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iG})$ be the RNAseq read counts training sample in the class label $c$ and $G$ is the total number of RNAseq.

The purpose of this study is to predict the class label $c$ of a future observation $Y_t$ using training samples associated with known class labels: $p(c|Y_t) \propto p(Y_t|c)p_c$, where $p_c$ is the probability of class $c$.

Using Bayes’ rule, we have,

$$p(c|Y_t) \propto p(Y_t|c)p_c;$$  \hspace{1cm} (1)

where, $p(Y_t|c)$ is the pdf of the sample $Y_t$ in class c, and $p_c$ is the prior probability that one sample comes from class $c$. The pdf of class-specific $c$ of RNAseq read counts of sample $Y_t$ and of RNAseq $g$ is,

$$P(Y_{ig} = y_{ig} | c) = \frac{\Gamma(\phi^{-1}_g + y_{ig})}{\Gamma(\phi^{-1}_g)} \left(\frac{\phi_g \mu_{igc}}{1 + \phi_g \mu_{igc}}\right)^{\phi^{-1}_g} \left(1 + \frac{1}{\phi_g \mu_{igc}}\right)^{y_{ig}}.$$  \hspace{1cm} (2)

In this parameterization, $Y_{ig}$ represents a count response of RNAseq, where $\mu_{igc}$ represents the mean, $\phi_g$ represents the dispersion parameter, $E(Y_{ig}) = \mu_{igc}$, and $Var(Y_{ig}) = \mu_{igc} + \mu_{igc} \phi^2_g$. Note we assume that all RNAseq are independent of each other, so we have,

$$p(Y_t|c) = \prod_{g=1}^{G} P(Y_{ig} = y_{ig}).$$  \hspace{1cm} (3)
Evaluating Equation (1) requires an estimation of \( p(Y_i|c) \) and \( p_e \). The model in Equation (2) states that \( Y_g \sim NB(\mu_{igc}, \phi_g) \). We first estimate \( \phi_1, \phi_2, \ldots, \phi_G, \) and \( \mu_{igc}, \mu_{igc}, \ldots, \mu_{igc} \) of all the training samples \( n_s \), and all RNAseq \( G \). The mean is estimated as
\[
\mu_{igc} = \bar{s}_igc,
\]
where \( s_{igc} \) is the size factor (26, 27) which is used to scale RNAseq counts for the \( ith \) sample (in class \( c \)). \( \lambda_{igc} \) is the total number of reads of RNAseq \( g \) across all samples in class \( c \). For prior \( p_e \), we assume all classes are equally likely, \( p_e = 1/C \). Note that \( \mu_{igc}, s_{igc}, \) and \( \lambda_{igc} \) are estimated for each class \( c \).

Next, plugging these estimates into Equation (2) and using the assumption of independent RNAseq, Equation (1) yields,
\[
\log(p(Y_i|c)) = \log(p(Y_i|c)) + \log(p_e).
\]
The log-likelihood log (\( p(Y_i|c) \)) is written as,
\[
\log(p(Y_i|c)) = \log\left(\prod_{g=1}^{G} P(Y_{ig} = y_{ig}|c)\right)
= \log\left(\prod_{g=1}^{G} \frac{\Gamma(\phi_g+1+y_{ig})}{\Gamma(\phi_g+1)\gamma_{ig}^{-1}} \exp\left(\frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\gamma_{ig}^{-1}\right)\right).
\]
Equation (5) can be written as,
\[
\log(p(Y_i|c)) = \sum_{g=1}^{G} \log\left(\frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\gamma_{ig}^{-1}\right)
+ \sum_{g=1}^{G} \log\left(\frac{1}{1+\phi_g\mu_{igc}}\phi_g^{-1}\right)
+ \sum_{g=1}^{G} \log\left(\frac{\Gamma(\phi_g+1+y_{ig})}{\Gamma(\phi_g+1)\gamma_{ig}^{-1}}\right).
\]
Rewriting Equation (6) yields,
\[
\log(p(Y_i|c)) = \sum_{g=1}^{G} \log\left(\frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\gamma_{ig}^{-1}\right)
- \sum_{g=1}^{G} \log\left(1+\phi_g\mu_{igc}\right)
+ \sum_{g=1}^{G} \log\left(\frac{\Gamma(\phi_g+1+y_{ig})}{\Gamma(\phi_g+1)\gamma_{ig}^{-1}}\right).
\]
The proposed NB model of genomics relates to the radiomics (imaging features) \( X \) through the mean parameters \( \mu_{igc} \) (estimated mean of an \( ith \) sample and RNAseq \( g \) in class \( c \)). We assume a log-linear regression model for estimating the mean \( \mu_{igc} \) in terms of the radiomics (imaging features) is given as follows:
\[
\log(\mu_{igc}) = X_i\beta_{igc}; \quad (8.a)
\log(s_{igc}) = X_i\beta_{igc}; \quad (8.b)
\]
where \( X_i \) is a \( p \)-dimensional of radiomics, \( \beta_{igc} \) is a \( p \)-dimensional vector of unknown regression coefficients (translate the relationship between \( X \) and \( Y \) through \( \mu_{igc} \)). The estimation of \( \beta_{igc} \) depends on class \( c \) and gene \( g \) of the \( ith \) sample. Hence, if there are two classes, we will need to estimate \( \beta_{i1} \) and \( \beta_{i2} \) (one from each class).

Plugging Equations (8.a) into Equation (7), yields,
\[
\log(p(Y_i|c)) = \log\left(\sum_{g=1}^{G} \frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\gamma_{ig}^{-1}\right)
+ \sum_{g=1}^{G} \log\left(\frac{1}{1+\phi_g\mu_{igc}}\phi_g^{-1}\right)
+ \sum_{g=1}^{G} \log\left(\frac{\Gamma(\phi_g+1+y_{ig})}{\Gamma(\phi_g+1)\gamma_{ig}^{-1}}\right).
\]
Using the estimated \( \beta_{igc} \), and \( \phi_g \) from the training data, we classify a test observation \( Y_i \) as follows,
\[
\log(p(Y_i|c)) = \log(p(Y_i|c)) + \log(p_e);
\]
and,
\[
\log(\log(p(Y_i|c))) = \log\left(\sum_{g=1}^{G} \frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\gamma_{ig}^{-1}\right)
+ \sum_{g=1}^{G} \log\left(\frac{1}{1+\phi_g\mu_{igc}}\phi_g^{-1}\right)
+ \sum_{g=1}^{G} \log\left(\frac{\Gamma(\phi_g+1+y_{ig})}{\Gamma(\phi_g+1)\gamma_{ig}^{-1}}\right) + \log(p_e).
\]

**Radiogenomics-NB Model Parameter Estimation**

**Estimating Dispersion \( \phi_g \) Using Weighted Likelihood Empirical Bayes**

Various methods for estimating the dispersion parameter are proposed in the literature. The EdgeR method applies a weighted conditional log-likelihood method to estimate the dispersion parameter (19). The weighted conditional log-likelihood (WL) for \( \phi_g \) is defined as a weighted combination of the individual (per-gene) likelihood \( l_g(\phi_g) \) and common \( l_c(\phi_g) \) likelihood:
\[
WL(\phi_g) = l_g(\phi_g) + \alpha l_c(\phi_g);
\]
where \( \alpha \) is the weight of \( l_c(\phi_g) \).

In EdgeR, \( \hat{\phi}_g \) is assumed to be normally distributed with means \( \phi_g \) and known variance \( \tau^2 \), and has the following hierarchical model:
\[
\hat{\phi}_g|\phi_g \sim N(\phi_g, \tau^2), \quad \text{and} \quad \phi_g | \phi_0 \sim N(\phi_0, \tau^2).
\]
Under this hierarchical normal model, the maximum weighted conditional log-likelihood estimator is given as:
\[
\hat{\phi}_g^{WL} = \frac{\hat{\phi}_g / \tau^2 + \alpha \sum_{g=1}^{G} \hat{\phi}_g / \tau^2}{1 / \tau^2 + \alpha \sum_{g=1}^{G} 1 / \tau^2};
\]
where,

$$1/\alpha = \sum_{i=1}^{G} \frac{\tau_i^2}{\tau_0^2}$$  \hspace{1cm} (15)$$

and,

$$\phi_0 = \hat{\phi}_0 = \frac{\sum_{i=1}^{G} \hat{\phi}_i / \tau_i^2}{1 / \tau_0^2}.$$  \hspace{1cm} (16)

**Computation of the Mean of RNAseq $\mu_{igc}$**

The size factor $s_{ic}$ of sample $i$ and class $c$ is the total number of RNAseq read counts of that sample divided by the total number of all RNAseq read counts across all training samples (in class $c$). The size factor estimation is vital to account for the different sequencing depth (library size) that may be used to sequence different samples and is computed as follows:

$$s_{ic} = \frac{\sum_{g=1}^{G} y_{igc}}{\sum_{i=1}^{n_c} \sum_{g=1}^{G} y_{igc}};$$  \hspace{1cm} (17)

where, $y_{igc}$ is the RNAseq read count of sample $i$ and RNAseq $g$ in class $c$, and $n_c$ is the total number of samples in class $c$.

The mean $\mu_{igc}$ of sample $i$ and RNAseq $g$ in class $c$ is then estimated as $\mu_{igc} = s_{ic}\lambda_{gc}$, where $\lambda_{gc}$ is the total number of reads per RNAseq in class $c$, and is computed as follows:

$$\lambda_{gc} = \sum_{i=1}^{n_c} y_{igc}.$$  \hspace{1cm} (18)

Using the estimated value of $\mu_{igc}$, the values of $\beta_{gc}$ are computed using equation 8.a as follows:

$$\beta_{gc} = x_i \log \left( \frac{\sum_{g=1}^{G} y_{igc} / \sum_{g=1}^{G} \sum_{i=1}^{n_c} y_{igc}}{N_c / \sum_{i=1}^{N_c} \sum_{g=1}^{G} y_{igc}} \right).$$  \hspace{1cm} (19)

The algorithm in **Figure 2** illustrates the steps of estimating the different parameters in the radiogenomics-NB classification model.

---

**EXPERIMENTAL RESULTS**

**Dataset**

The dataset in this study consists of 108 pre-operative lower grade glioma (LGG) patients that are described in Menze et al. (28), Bakas et al. (29), and Akas et al. (30). Four sequences of the MRI are provided with the dataset: pre-contrast T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). These scans are skull-stripped, re-sampled to $1 \text{mm}^3$ resolution, and co-registered to the T1 template. The dataset provides the segmented sub-regions of the LGG: Gadolinium enhancing tumor (ET), the peritumoral edema (ED), and necrosis along with non-enhancing tumor (NCR/NET).

RNAseq read counts data (with a total number of 56830 RNAseq), molecular alterations (IDH mutation, 1p/19q codeletion, and ATRX), grade (II and III), and the clinical dataset can be found and downloaded from The Cancer Genome Atlas (TCGA) dataset in the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). RNAseq are primarily obtained from solid portions of tumor. The clinical dataset is de-identified in compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The distribution of the data is as follows: (i) IDH mutation: 85 Mutant and 23 wildtype (WT), (ii) 1p/19q codeletion: 27 codeletion and 81 non-codeletion, and (iii) ATRX status: 43 Mutant and 65 WT. The range of the patients’ age at diagnosis is 20–75 years, and the median age is 46.5 years.

**Data Preparation**

In this study we first filter RNAseq read counts to remove RNAseq with very low value of read counts before performing any statistical analysis. RNAseq with very low read counts hold very little information because an RNAseq of biological importance needs to be expressed at some minimal level. We utilize a quantile filter (31) with a quantile threshold of 0.25. This step returns each RNAseq that has a mean across all samples higher than the defined quantile threshold of 0.25. Then, we reduce the number of RNAseq that are used in the radiogenomics-NB models, by...
TABLE 1 | Radiomics features description and their ANOVA p-value association with IDH mutations, 1p/19q codeletion, and ATRX mutations.

| Feature number | Feature description                                      | p-value of IDH mutation | p-value of 1p/19q codeletion | p-value of ATRX mutation |
|----------------|----------------------------------------------------------|-------------------------|------------------------------|-------------------------|
| 1              | the size of the enhancing tumor to the necrosis size     | <0.005                  | 0.393                        | 0.178                   |
| 2              | the size of the enhancing tumor to the size of enhancing tumor and necrosis | 0.8630                  | 0.070                        | 0.239                   |
| 3              | the size of the enhancing tumor to the edema size        | <0.005                  | 0.600                        | <0.005                  |
| 4              | the size of the enhancing tumor to the whole tumor size  | <0.005                  | 0.707                        | 0.027                   |
| 5              | the size of the edema to the necrosis size               | 0.188                   | 0.996                        | 0.114                   |
| 6              | the size of the edema to the size of enhancing tumor and necrosis | 0.138                   | 0.789                        | 0.0237                  |
| 7              | the size of the edema to the whole tumor size            | <0.005                  | 0.131                        | <0.005                  |
| 8              | and the size of the necrosis to the whole tumor size     | <0.005                  | 0.221                        | <0.005                  |

utilizing EdgeR (18–20) to extract the differentially expressed RNAs (DEGs). DEGs reflect the significance of a gene in a certain biological condition. In this study, we select the top 10, 20, 30, 50, 100, and 150 DEGs (see Supplementary Table 1).

Furthermore, we use eight volumetric radiomics features as illustrated in Table 1. ANOVA analysis for radiomics in Table 1 shows that feature numbers 1, 3, 4, 7, and 8 are significantly associated (ANOVA test, p < 0.05) with IDH mutations as illustrated in Figure 3A. Our analysis also indicates that feature number 2 is marginally associated (ANOVA test, p = 0.07) with 1p/19q codeletion. Furthermore, our analysis indicates that feature numbers 3, 4, 6, 7, and 8 are significantly associated (ANOVA test, p < 0.05) with ATRX mutations as illustrated in Figure 3B. Additionally, our analysis reveals that thresholding feature number 6 around the mean creates an ordinal feature that is significantly associated (ANOVA test, p < 0.05) with 1p/19q codeletion, and ATRX mutations. Likewise, thresholding feature numbers 1, 3, 5, 7, and 8 around their means converts these features into ordinal features that are significantly associated (ANOVA test, p < 0.05) with IDH and ATRX mutations. Moreover, thresholding feature numbers 5, 6, 7, and 8 around their median converts these features into ordinal features that are significantly associated (ANOVA test, p < 0.05) with IDH and ATRX mutations.

Few other studies suggest that these volumetric imaging features and their ratios are associated with and predictive of several mutations in gliomas (32–35).

The 108 LGG cases are randomly split into 80% training and 20% testing sets, and a balanced distribution of the target molecular alteration is ensured in the training and testing sets in each molecular classifier. The trained model classifier is developed using the training set. Model performance prediction is estimated and reported using the testing sets in terms of accuracy, balanced accuracy, F1 score, sensitivity, specificity, negative predictive value, and positive predictive value. The training set is utilized to build our radiogenomics-NB classifier as shown in steps 1-4 in Figure 2. The testing set is used to estimate the performance of the classifier as shown in steps a and b in Figure 2. Authors In Dong et al. (23), Maufroy et al. (36), Pan et al. (37), and Vabalas et al. (38) repeat training and testing analysis for a specific number of times to ensure the robustness of the model performance. Consequently, in this work, we repeat the whole procedure 100 times independently for the 3 molecular alterations and then report the mean and standard deviation of the classifiers’ performance using the testing sets.

Model performance parameters are computed based on the confusion matrix in Figure 4 as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN};
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN};
\]

\[
\text{Specificity} = \frac{TN}{TN + FP};
\]

\[
\text{Positive predictive value} = \frac{TP}{TP + FP};
\]

\[
\text{Negative predictive value} = \frac{TN}{TN + FN};
\]

\[
\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}, \text{ and (25)}
\]

\[
\text{F1 score} = \frac{TP}{(TP + \frac{FP + FN}{2})}; \text{ (26)}
\]

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the true negative.

**Radiogenomics-NB Models Using Different Number of Differently Expressed RNAs**

In this section, we investigate the importance of using different numbers of DEGs on the performance of the radiogenomics-NB model. LGG radiogenomics-NB mutation prediction models are developed based on the top 10, 20, 30, 50, 100, and 150 DEGs. The performance of the radiogenomics-NB IDH model using the top 10 DEGs achieves slightly higher performance. However, such improvement is not statistically significant (ANOVA test, p > 0.05) when compared to the performance of the IDH models with the other number of DEGs (Figure 5A) except for negative predictive value (NPV) performance when using the top 20 DEGs. Using the top 20 DEGs in the IDH model achieves significantly worse NPV when compared to the NPV achieved using the top 10 DEGs (ANOVA test, p < 0.05). Radiogenomics-NB IDH model with the top 10 DEGs (red line in Figure 5A) achieves an overall accuracy (Acc) of 0.92 ± 0.06, sensitivity (Sens) of 0.94 ± 0.07, specificity (Spec)
of 0.83 ± 0.18, positive predictive value (PPV) of 0.96 ± 0.04, negative predictive value (NPV) of 0.82 ± 0.17, F1 score of 0.95 ± 0.04, and balanced accuracy (B. Acc) of 0.88 ± 0.09, respectively.

Radiogenomics-NB codeletion models achieve similar performance (ANOVA test, $p > 0.05$) using the top 10, 20, 30, and 50 DERs as shown in Figure 5B. Furthermore, using the top 100 and 150 DERs in the codeletion model achieves significantly
worse performance when compared to the performance of using the top 10 DERs (ANOVA test, $p < 0.05$). Using the top 10 DERs, the radiogenomics-NB codelletion model achieves an accuracy of 0.93 ± 0.06, a balanced accuracy of 0.90 ± 0.10, F1 score of 0.86 ± 0.14, a sensitivity of 0.84 ± 0.19, a specificity of 0.96 ± 0.04, an NPV of 0.95 ± 0.06, and a PPV of 0.90 ± 0.12, respectively.

Radiogenomics-NB ATRX model also achieves similar performance (ANOVA test, $p > 0.05$) using the top 10, 20, and 30 DERs, even though the performance when using the top 10 DERs is slightly better as illustrated in Figure 5C. Using the top 10 DERs, the ATRX model achieves an accuracy of 0.85 ± 0.07, a balanced accuracy of 0.85 ± 0.07, an F1 score of 0.82 ± 0.08, a sensitivity of 0.86 ± 0.13, a specificity of 0.85 ± 0.09, an NPV of 0.91 ± 0.08, and a PPV of 0.80 ± 0.10, respectively.

**Comparative Analysis**

Figure 6 illustrates a graphical performance comparison between our radiogenomics-NB model with that of four different classifiers in the literature: NBLDA (23), VoomNSC (12, 13), RF-genomics where we first log-transformed (20) the RNAseq into a normal distribution, and RF-radiogenomics. Note that the number of DERs that we apply to develop these classifiers is 10 DERs. Moreover, when developing these classifiers, the 108 LGG cases are randomly split into 80% training and 20% testing sets, and balanced distribution is ensured when developing the different classifiers. The trained model classifier is developed using the training set, and 10-fold cross-validation is performed to identify the tuning parameters in the different classifiers. Model performance prediction is estimated and reported using the testing sets. Additionally, to ensure the robustness of the different classifiers’ performance, we repeat the whole procedure 100 times independently and every training/testing set is utilized to develop and estimate the performance of each classifier.

The NBLDA (23) classifier is developed by fitting NB to the top 10 DERs; then the mean and dispersion parameter are estimated from these DERs. In RF-genomics, the top 10 DERs of the training sets are first log-transformed into normal distribution and then fed into RF to build the RF-genomics classifier. In RF-radiogenomics, radiomics (eight volumetric features described previously in section Data Preparation) are utilized with the log-transformed DERs and then fed into RF to build the RF-radiogenomics classifier. VoomNSC (12, 24) is developed by first applying the Voom-based transformation on the 10 DERs and then applying the NSC classifier as illustrated in Zararsiz et al. (12) and Tibshirani et al. (24).

Comparing the performance of our radiogenomics-NB IDH model with that of NBLDA, RF-genomics, and VoomNSC, the radiogenomics-NB IDH significantly outperforms (ANOVA test, $p < 0.05$) these methods as shown in Figure 6A and Table 2. Additionally, our radiogenomics-NB IDH model significantly
Our radiogenomics-NB codeletion model (Figure 6B and Table 3) performance is similar to NBLDA, RF-genomics, VoomNSC, and RF-radiogenomics models, except for the specificity and NPV performance when using RF-radiogenomics and VoomNSC. The specificity and NPV of our model are significantly higher than those achieved by RF-genomics and VoomNSC. Our radiogenomics-NB codeletion model achieves an accuracy of 0.93 ± 0.06, a sensitivity of 0.84 ± 0.20, a specificity of 0.96 ± 0.5, an F1 score of 0.86 ± 0.14, and a balanced accuracy of 0.90 ± 0.10, respectively.

The performance of our radiogenomics-NB ATRX model as shown in Figure 6C and Table 4 outperforms both NBLDA and VoomNSC significantly (ANOVA test, \( p < 0.05 \)). However, comparing our ATRX model to RF-genomics, our model achieves significantly better balanced-accuracy, F1 score, NPV, and sensitivity. Additionally, comparing our ATRX model to RF-radiogenomics, our model achieves significantly (ANOVA test, \( p < 0.05 \)) better sensitivity but achieves similar accuracy, balanced-accuracy, F1 score, and sensitivity. Our radiogenomics-NB ATRX model achieves an accuracy of 0.85 ± 0.07, a sensitivity of 0.86 ± 0.13, a specificity of 0.85 ± 0.09, an F1 score of 0.82 ± 0.08, and a balanced accuracy of 0.85 ± 0.07, respectively. The RF-radiogenomics ATRX model achieves an accuracy of 0.84 ± 0.08, a sensitivity of 0.80 ± 0.14, a specificity of 0.86 ± 0.10, an F1 score of 0.80 ± 0.09, and a balanced accuracy of 0.83 ± 0.08, respectively.

**Gender-Specific Effect Analysis of Radiogenomics-NB**

In our LGG dataset, IDH mutated patients, unlike IDH WT patients, have significantly longer survival (65.7 vs. 19.9 months, log-rank test \( p = 0.004 \)). The association between IDH status and overall survival remains significant after stratifying for gender (likelihood ratio test \( p = 0.015 \)). However, the association between 1p/19q codeletion and ATRX status and overall survival is not significant. Additionally, the chi-square test shows no significant association \( (p > 0.05) \) between gender and IDH status, 1p/19q co-deletion, and ATRX status. Table 5 shows patient IDH status, 1p/19q codeletion, and ATRX status distribution based on gender.

To explore the gender-specific effect in the performance of the radiogenomics-NB, we build two radiogenomics-NB models based on gender; male-specific radiogenomics-NB and female-specific radiogenomics-NB. Our analysis indicates that female-specific models significantly outperform (ANOVA test, \( p < 0.05 \)) male-specific models as illustrated in Figure 7. In the radiogenomics-NB IDH, female-specific model achieves an accuracy of 0.93 ± 0.08, a sensitivity of 0.93 ± 0.09, a specificity of 0.91 ± 0.10, a PPV of 0.97 ± 0.05, an NPV of 0.83 ± 0.21, and a balanced accuracy of 0.92 ± 0.11, respectively. The male specific IDH model achieves an accuracy of 0.85 ± 0.08, a sensitivity of 0.97 ± 0.06, a specificity of 0.35 ± 0.33, a PPV of 0.86 ± 0.07, an NPV of 0.55 ± 0.48, and a balanced accuracy of 0.66 ± 0.17, respectively.

In the radiogenomics-NB codeletion, female-specific model achieves an accuracy of 0.91 ± 0.09, a sensitivity of 0.77 ±
**TABLE 2** | Probability of significant difference using ANOVA test between the differentially expressed radiogenomics-NB model and different classifiers using the IDH dataset.

| IDH         | Accuracy | Sensitivity | Specificity | PPV  | NPV  | F1    | Balanced accuracy |
|-------------|----------|-------------|-------------|------|------|-------|-------------------|
| radiogenomics-NB vs. NBLDA | 0.000    | 0.010       | 0.000       | 0.000| 0.000| 0.000 | 0.000             |
| radiogenomics-NB vs. VoomNSC  | 0.000    | 0.075       | 0.000       | 0.000| 0.000| 0.000 | 0.000             |
| radiogenomics-NB vs. RF       | 0.001    | 0.023       | 0.000       | 0.000| 0.138| 0.000 | 0.000             |
| radiogenomics-NB vs. RF-radiogenomics | 0.069 | 0.432       | 0.061       | 0.000| 0.084| 0.001 | 0.01              |

A statistically significant difference exists if \( p < 0.05 \). Values in bold show a significant improvement of our radiogenomics-NB IDH over the compared one.

**TABLE 3** | Probability of significant difference using ANOVA test between the differentially expressed radiogenomics-NB model and different models using the 1p/19q codeletion dataset.

| CODEL       | Accuracy | Sensitivity | Specificity | PPV  | NPV  | F1    | Balanced accuracy |
|-------------|----------|-------------|-------------|------|------|-------|-------------------|
| radiogenomics-NB vs. NBLDA | 0.232    | 0.186       | 0.756       | 0.514| 0.253| 0.123 | 0.181             |
| radiogenomics-NB vs. VoomNSC  | 0.072    | 0.228       | 0.001       | 0.057| 0.042| 0.742 | 0.317             |
| radiogenomics-NB vs. RF       | 0.242    | 0.390       | 0.020       | 0.636| 0.027| 0.271 | 0.42              |
| radiogenomics-NB vs. RF-radiogenomics | 0.671 | 0.815       | 0.893       | 0.825| 0.282| 0.855 | 0.792             |

A statistically significant difference exists if \( p < 0.05 \). Values in bold show a significant improvement of our radiogenomics-NB codeletion over the compared one.

**TABLE 4** | Probability of significant difference using ANOVA test between the differentially expressed radiogenomics-NB model and different models using the ATRX dataset.

| ATRX        | Accuracy | Sensitivity | Specificity | PPV  | NPV  | F1    | Balanced accuracy |
|-------------|----------|-------------|-------------|------|------|-------|-------------------|
| Radiogenomics-NB vs. NBLDA | 0.000    | 0.269       | 0.000       | 0.000| 0.677| 0.004 | 0.001             |
| Radiogenomics-NB vs. VoomNSC  | 0.003    | 0.741       | 0.001       | 0.002| 0.432| 0.021 | 0.012             |
| Radiogenomics-NB vs. RF       | 0.083    | 0.005       | 0.540       | 0.960| 0.004| 0.026 | 0.025             |
| Radiogenomics-NB vs. RF-radiogenomics | 0.183 | 0.003       | 0.215       | 0.561| 0.003| 0.052 | 0.053             |

A statistically significant difference exists if \( p < 0.05 \). Values in bold show a significant improvement of our radiogenomics-NB ATRX over the compared one.

**TABLE 5** | Gender-based distribution of IDH status, 1p/19q codeletion, and ATRX status in the LGG dataset.

| IDH status | Mutant | WT | 1p/19q codeletion | Non-codeletion | ATRX status | Mutant | WT |
|------------|--------|----|------------------|----------------|-------------|--------|----|
| Mutant     | 43     | 14 | 14               | 14             | 24          | 33     |
| Male       | 42     | 9  | 13               | 38             | 19          | 32     |

0.31, a specificity of 0.96 ± 0.07, a PPV of 0.80 ± 0.32, an NPV of 0.93 ± 0.07, and a balanced accuracy of 0.84 ± 0.17, respectively. The male specific codeletion model achieves an accuracy of 0.84 ± 0.10, a sensitivity of 0.56 ± 0.32, a specificity of 0.95 ± 0.08, a PPV of 0.79 ± 0.32, an NPV of 0.86 ± 0.10, and a balanced accuracy of 0.77 ± 0.17, respectively.

In the radiogenomics-NB ATRX, female-specific model achieves an accuracy of 0.80 ± 0.11, a sensitivity of 0.79 ± 0.20, a specificity of 0.81 ± 0.15, a PPV of 0.76 ± 0.16, an NPV of 0.87 ± 0.12, and a balanced accuracy of 0.80 ± 0.12, respectively. The male specific ATRX model achieves an accuracy of 0.76 ± 0.12, a sensitivity of 0.69 ± 0.23, a specificity of 0.81 ± 0.14, a PPV of 0.73 ± 0.18, an NPV of 0.81 ± 0.12, and a balanced accuracy of 0.75 ± 0.13, respectively.

**DISCUSSION**

In this study, we propose a novel radiogenomics-NB model to fuse radiomics (imaging features) with RNAseq (genes) for glioma grading and prediction. NB distribution is appropriate for modeling RNAseq discrete read counts data and for preserving the count-based nature of this data. In the proposed radiogenomics-NB model, log-linear regression modeling is fitted to the estimated mean of the NB distribution and is linked with radiomics. We introduce this step to fuse the continuous radiomics data with the RNAseq count-based data without the need to transform the RNAseq data into a normal distribution.

The NB, unlike a Poisson distribution, has two parameters; the mean (e.g., the expected value of the RNAseq read counts data) and dispersion (e.g., a parameter that helps in capturing
the variability of the RNAseq read counts). If the dispersion of NB is zero, the model reduces to Poisson distribution. In Poisson distribution, the mean is equal to the variance, which makes it rather restrictive. However, variation is usually observed in the real data of RNAseq counts data that the Poisson distribution cannot handle properly. On the other hand, NB has an additional parameter called the “dispersion” that allows the NB distribution of RNAseq counts data to modify its variance without affecting the mean. Thus, NB serves as a practical approximation to model RNAseq count data with variability different from its mean.

The mean of the proposed radiogenomics-NB model is estimated as the size factor multiplied by the total number of reads per RNAseq. Moreover, we utilize EdgeR to estimate the dispersion of the proposed radiogenomics-NB assuming RNAseq variability is assessed using the weighted conditional log-likelihood model. In the weighted conditional model, RNAseq counts data is assumed to have a distinct and individual dispersion for each RNAseq in addition to a common dispersion. Such an assumption can be more reliable when estimating the dispersion of real data of RNAseq counts data.

The performance evaluation of the proposed work indicates that linking simple, clinically feasible radiomics (i.e., tumor volumetric features) to RNAseq improves the performance of IDH and ATRX mutations prediction. The radiomics features utilized in the proposed radiogenomics-NB model that are described in Table 1 mainly depend on volumetric features. Our analysis shows that these features are associated with particular glioma mutations. This outcome supports previous studies that show the association between volumetric features and glioma mutations (32–35). The efficacy of the proposed radiogenomics-NB model is further investigated using the top 10, 20, 30, 50, 100, and 150 DERs, respectively. Our analysis shows that the smaller the number of DERs (fewer than 30 DERs) utilized in radiogenomics-NB, the better is the radiogenomics-NB model performance. Our analyses indicate that using fewer than 30 DERs in our analysis offers the best performance (statistically significant) in the radiogenomics-NB codeletion and ATRX prediction model. This suggests that using large numbers of DERs (more than 30) in the proposed radiogenomics-NB would over parametrize the dataset and create model fitting problems and thus degrade the performance.

Comparing our radiogenomics-NB model to NBLDA, RF-genomics, FR-radiogenomics, and VoomNSC, our model significantly outperforms NBLDA, RF-genomics, and VoomNSC for prediction of IDH and ATRX mutations. Our radiogenomics-NB model offers similar performance as NBLDA, RF-genomics, and VoomNSC models for prediction of 1p/19q codeletion. Specifically, for prediction of IDH mutations, while the proposed radiogenomics-NB model achieves significantly better balanced-accuracy, F1 score, and PPV than RF-radiogenomics, our model achieves similar accuracy, sensitivity, and specificity. Such results indicate the power of fusing radiomics and genomics data to develop radiogenomics models for classification and prediction models. The findings in this work indicate that the radiomics volumetric features may be vital for the prediction of IDH and ATRX mutations along with the genomics.

Different studies have revealed that gender is a significant factor in identifying cancer survival, prognosis, and treatment response (39–41). Hence, improved glioma molecular mutation prediction may require the development of gender-specific models. In this study, we explore the gender-specific effect on the radiogenomics-NB models. Our analysis reveals that IDH mutated patients remain significant after stratifying for gender, unlike 1p/19q codeletion and ATRX status. Moreover, our analysis

**FIGURE 7** | Gender-based radiogenomics-NB models performance. (A) IDH mutations, (B) 1p/19q codeletion, and (C) ATRX mutations which are computed across 100 testing sets. The error bar represents one standard deviation. The asterisk * illustrates a significant difference between the two measurements. Y-axis represents the average performance of the different statistics on the X-axis. Different colors represent the female- and male-specific radiogenomics-NB models.
indicates that no association is found between gender and the three specific mutations (IDH mutations, 1p/19q codeletion, and ATRX status) using the Chi-square test. This result is in agreement with the findings in Brat et al. (42), Li et al. (43), and Ebrahimi et al. (44). However, our gender-specific modeling shows that female-specific radiogenomics-NB models significantly outperform the male-specific radiogenomics-NB models for prediction of IDH status, 1p/19q codeletion, and ATRX status, respectively.

In conclusion, we present a glioma mutations radiogenomics-NB prediction model that preserves the count nature of RNAseq counts data in the NB model and utilizes radiomics to develop a complete and a better characterization prediction model of patient data. Our analysis shows the superiority of utilizing both genomics and clinically feasible radiomics data when compared to only genomics models. Use of tumor volumetrics can be more easily and reproducibly implemented in clinical practice compared to more complex radiomics metrics, such as higher order texture analysis features. Finally, this study can be more easily and reproducibly implemented in clinical practice compared to only genomics models. Use of tumor volumetrics can be more easily and reproducibly implemented in clinical practice compared to more complex radiomics metrics, such as higher order texture analysis features. Furthermore, it will be interesting to investigate the cause of superior performance of female-specific radiogenomics-NB models when compared to that of the male-specific radiogenomics-NB models for prediction of IDH status, 1p/19q codeletion, and ATRX status. Also, these models may be further investigated in treatment response and survival prediction in the future.

DATA AVAILABILITY STATEMENT
Radiomics data are available at doi: 10.7937/K9/TCIA.2017. GJQ7R0EF. Genomics and mutations data are available at: https://portal.gdc.cancer.gov/.

ETHICS STATEMENT
The studies involving human participants were reviewed and approved by Old Dominion University IRB. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS
ZS and KI: conception and design and development of methodology. ZS, ND, and KI: analysis and interpretation of data. ZS, ND, AV, JC, and KI: drafting the article and/or revising. KI: funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING
This work is supported by National Institutes of Health (R01 EB020683).

SUPPLEMENTARY MATERIAL
The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.705071/full#supplementary-material

REFERENCES
1. Weisleder R, Schaewiger MC, Gambhir SS, Hricak H. Imaging approaches to optimize molecular therapies. Sci Transl. Med. (2016) 8:355ps316–355ps316. doi: 10.1126/scitranslmed.aaf3936
2. O’Connor JP, Abagge EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. (2017) 14:169. doi: 10.1038/nrclinonc.2016.162
3. Reza SM, Samad MD, Shboul ZA, Jones KA, Iftekharuddin KM. Glioma grading using structural magnetic resonance imaging and molecular data. J Med Imaging. (2019) 6:024501. doi: 10.1117/1.JMI.6.2.024501
4. Shboul ZA, Iftekharuddin KM. Prediction of low-grade glioma progression using MR imaging. In: Medical Imaging 2019: Computer-Aided Diagnosis: International Society for Optics and Photonics. San Diego, CA (2019).
5. Shboul ZA, Iftekharuddin KM. Efficacy of radiomics and genomics in predicting TP53 mutations in diffuse lower grade glioma. In: Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging: International Society for Optics and Photonics. Houston, TX (2020).
6. Kickingereder P, Bonekamp D, Nowosieslki M, Kritz A, Sill M, Burth S, et al. Radiogenomics of glioblastoma: machine learning-based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. Radiology. (2016) 281:907–18. doi: 10.1148/radiol.2016161382
7. Mazurowski MA, Clark K, Czarneck NM, Shamesfandabadi P, Peters KB, Saha A. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data. J Neurooncol. (2017) 133:27–35. doi: 10.1007/s11060-017-2420-1
8. Rathore S, Akbari H, Rozyczki M, Abdullah KG, Nasrallah MP, Binder ZA, et al. Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. Sci Rep. (2018) 8:5087. doi: 10.1038/s41598-018-2739-2
9. Iwadate Y, Sakaïda T, Hiwasa T, Nagai Y, Ishikura H, Takiguchi M, et al. Molecular classification and survival prediction in human gliomas based on proteome analysis. Cancer Res. (2004) 64:2496–501. doi: 10.1158/0008-5472.CAN-03-1254
10. Zhang X, Sun S, Pu JKS, Tsang ACO, Lee D, Man VOY, et al. Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol Dis. (2012) 48:1–8. doi: 10.1016/j.nbd.2012.06.004
11. Zeng W-J, Yang Y-L, Liu Z Z, Wen Z-P, Chen Y-H, Hu X-L, et al. Integrative analysis of DNA methylation and gene expression identify a three-gene signature for predicting prognosis in lower-grade gliomas. Cellular Physiology and Biochemistry. (2018) 47:428–39. doi: 10.1159/000489954
12. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. (2014) 15:R29. doi: 10.1186/gb-2014-15-2-r29
13. Zarariz G, Goksuluk D, Klaus B, Korkmaz S, Eldem V, Karabulut E, et al. voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. PeerJ. (2017) 5:e3890. doi: 10.7717/peerj.3890
14. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. (2010) 11:1–10. doi: 10.1186/gb-2010-11-12-220

15. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. (2012) 40:4288–97. doi: 10.1093/nar/gks042

16. Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: poisson, overdispersed Poisson, and negative binomial models. Psychol Bull. (1995) 118:392. doi: 10.1037/0033-2909.118.3.392

17. Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harbor Protoc. (2015) 2015:pdb.top084970. doi: 10.1101/pdb.top084970

18. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. (2007) 9:321–32. doi: 10.1093/biostatistics/kxm030

19. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. (2007) 23:2881–7. doi: 10.1093/bioinformatics/btm453

20. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616

21. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. (2010) 11:R106. doi: 10.1186/gb-2010-11-10-R106

22. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat Appl Genet Mol Biol. (2011) 10:1–28. doi: 10.2202/1544-6115.1637

23. Dong K, Zhao H, Tong T, Wan X. NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. BMC Bioinformatics. (2016) 17:369. doi: 10.1186/s12859-016-1208-1

24. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science. (2003) 18:104–17. doi: 10.1214/ss/1056394888

25. Louis DN, Perry A, Reifenberger G, Von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. (2016) 131:803–20. doi: 10.1007/s00401-016-1545-1

26. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. (2008) 18:1509–17. doi: 10.1101/gr.079558.108

27. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. (2008) 5:621. doi: 10.1038/nmeth.1226

28. Menzle BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. (2015) 34:1993–2024. doi: 10.1109/TMI.2014.2377694

29. Bakas S, Akbari H, Sotiras A, Béchet J, Rovosky M, Kirby J, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nat Sci Data. (2017) 4:170117. doi: 10.1038/sdata.2017.117

30. Bakas S, Akbari H, Sotiras A, Béchet J, Rovosky M, Kirby J, et al. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. Cancer Imaging Arch. (2017) 286. doi: 10.7977/KVTCIA.2017.GJQR0EF

31. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. (2015) 44:e171. doi: 10.1093/nar/gkuv1507

32. Mettellus P, Coulbaly B, Colin C, De Paula AM, Vasiljevic A, Taieb D, et al. Absence of IDH mutation identifies a novel radiologic and molecular subtype of WHO grade II gliomas with dismal prognosis. Acta Neuropathol. (2010) 120:719–29. doi: 10.1007/s00401-010-0777-8

33. Gutman DA, Dunn WD, Grossmann P, Cooper LA, Holder CA, Ligon KL, et al. Somatic mutations associated with MRI-derived volumetric features in glioblastoma. Neuroradiology. (2015) 57:1227–37. doi: 10.1007/s00234-015-1576-7

34. Park Y, Han K, Ahn S, Bae S, Choi Y, Chang J, et al. Prediction of IDH1-mutation and 1p/19q-codelesion status using preoperative MR imaging phenotypes in lower grade gliomas. Am J Neuroradiol. (2018) 39:37–42. doi: 10.3174/ajnr.A5421

35. Thust S, Hassanein S, Bisdas S, Rees J, Hyare H, Maynard J, et al. Apparent diffusion coefficient for molecular subtyping of non-gadolinium-enhancing WHO grade II/III glioma: volumetric segmentation versus two-dimensional region of interest analysis. Eur Radiol. (2018) 28:3779–88. doi: 10.1007/s00330-018-5351-0

36. Mauratoy A, Chassot E, Joo R, Kaplan DM. Large-scale examination of spatio-temporal patterns of drifting fish aggregating devices (DFADs) from tropical tuna fisheries of the Indian and Atlantic Oceans. PLoS ONE. (2015) 10:e0128203. doi: 10.1371/journal.pone.0128203

37. Pan L, Liu G, Lin E, Zhong S, Xia H, Sun X, et al. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci Rep. (2017) 7:1–9. doi: 10.1038/s41598-017-07408-0

38. Vabulas AR, Goven E, Poliakov E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS ONE. (2019) 14:e0224365. doi: 10.1371/journal.pone.0224365

39. Yuan Y, Liu L, Chen H, Wang Y, Xu Y, Mao H, et al. Comprehensive characterization of molecular differences in cancer between male and female patients. Cancer Cell. (2016) 29:711–22. doi: 10.1016/j.ccell.2016.04.001

40. Ippolito E, Yim AK-Y, Luo I, Chinnaiyan P, Rubin J. Sexual dimorphism in glioma glycolysis underlies sex differences in survival. JCI Insight. (2017) 2:92142. doi: 10.1172/jci.insight.92142

41. Yang W, Warrington NM, Taylor SJ, Whitmire P, Carrasco E, Singleton KW, et al. Sex differences in GBM revealed by analysis of patient imaging, transcriptome, and survival data. Sci Transl Med. (2019) 11:eaa0.5253. doi: 10.1126/scitranslmed.aao5253

42. Drat DJ, Verhaak RG, Aldape KD, Yung WA, Salama SR, Cooper LA, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New Engl J Med. (2015) 372:2481–98. doi: 10.1056/NEJMoa1420121

43. Li M-Y, Wang Y-Y, Cai J-Q, Zhang C-B, Wang K-Y, Cheng W, et al. Isocitrate dehydrogenase I gene mutation is associated with prognosis in clinical low-grade gliomas. PLoS ONE. (2015) 10:e0130872. doi: 10.1371/journal.pone.0130872

44. Ebrahimi A, Skardelly M, Bonzhem I, Ott I, Mühleisen H, Eckert I, et al. ATRX immunostaining predicts IDH and H3F3A status in gliomas. Acta Neuropathol Commun. (2016) 4:600. doi: 10.1186/s40478-016-0331-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Shboul, Diawara, Vossough, Chen and Itkonenraddin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.