Multi-path convolutional neural network for glioblastoma survival group prediction with point mutations and demographic features

1st Abdulrhman Aljouie*
Department of Biostatistics and Bioinformatics
King Abdullah International Medical Research Center
Riyadh, Saudi Arabia
aljouieab@ngha.med.sa

2nd Usman Roshan†
*Department of Computer Science**†
New Jersey Institute of Technology
Newark, New Jersey
aa547@njit.edu*, usman@njit.edu†

***Abstract*—Glioblastoma multiforme (GBM) is the most common and aggressive brain cancer with a median survival rate of 15 months. It is well-established that age is a strong independent predictor of GBM survival outcome. There is accumulating evidence that single nucleotide polymorphisms (SNPs) in the IDH1 gene influences GBM survival time. We propose a new multi-path convolutional neural network that combines SNPs, age, age groups, and gender to predict survival groups with a one-year threshold. We obtained GBM SNP and demographic data from The Cancer Genome Atlas. We compare our multi-path CNN with a support vector machine (SVM) and random forest. We randomly held out 10% of the samples as a test set, and employed 10-fold cross-validation for hyperparameter tuning in the remaining 90%. We then fit a model with optimal hyperparameters and predict the test set. In the combined SNP and demographic features, our proposed multi-path model achieved 67% accuracy in the test set compared to SVM accuracy of 60% and random forest accuracy of 47%. In the 10-fold cross-validation, our model predicted the two survival groups with 63% mean balanced accuracy while SVM and random forest attained 56% and 49% mean balanced accuracy. We evaluated the predictive ability in combined SNP and demographic data versus each data source alone for our proposed CNN, SVM, and random forest. The highest achieved accuracy for SNP data only in the test data set is 60% with our single-path CNN. The top accuracy in the test data set for demographic features alone attained is 60% by SVM and our single-path neural network.**

***Index Terms*—TCGA-GBM, survival prediction, convolutional neural network, SNP**

I. Introduction

Glioblastoma multiforme (GBM) is the most common and aggressive type of brain cancer, with a median survival rate of less than one and a half years [1]. Untreated patients with GBM have a median survival time of 3 months [2]. It is well-established that age is a strong independent predictor of survival time in gliomas [3]–[5]. Several studies have found that gender is significantly correlated [6]–[8]. A study that analyzed 6586 GBM patients shows that age and gender, among other seven features, are independent survival prognostic factors [9]. Other studies investigated the role of Single Nucleotide Polymorphisms (SNPs) on GBM overall survival outcomes [10], [11]. One study found that GBM patients who carry both TERT mutations and homozygous C-allele mutation for SNP rs2853669 have shorter survival time versus patients with wild-type allele [12]. There is accumulating evidence in the literature that GBM patients with IDH1 somatic mutation have significantly higher overall survival time compared to patients who carry a wild-type allele [13]–[15].

We hypothesize that combining tumor sample's SNP, age, and gender data increase the predictive power of GBM survival outcome. We propose a multi-path neural network to predict short ($<$ one-year) and long ($\geq$ one-year) survival groups. We assessed the predictive ability of combined SNPs, demographic features (age, age groups, and gender) versus each data source alone, and compared our method to support vector machine (SVM) with linear kernel, and random forest classifiers.

We downloaded The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) of 272 white individuals demographics (age, gender), survival (days from diagnosis to death), and tumor samples' pre-aligned whole-exome sequencing data from National Cancer Institute's Genomic Data Commons (GDC) portal. To obtain SNP data from sequence alignment files, we performed variant calling with Genome Analysis Toolkit (GATK, version 3.8) [16] followed by two-layers quality controls: 1) variant quality score recalibration (VQSR), and 2) hard filtering (depth < 5, genotype quality < 20). We excluded SNPs that have any missing value from further analysis.

We randomly held out 10% of the whole data set, 5% from each class to create a balanced subset, and and kept it as test set. We used the other 90% of data for training and hyperparameters tuning, by employing 10-fold cross-validation. We then fit a model with the 90% of data that is kept for training with best-performing hyperparameters and predict the test data set. We report the accuracy of SNPs alone, age and gender alone, and combined SNPs, age, and gender. We then compare the performance of our proposed method to SVM and random forest.

On the test dataset, the best classification performance is reached by feeding SNP and demographic features into our proposed multi-path convolutional neural network. There we achieved an accuracy of 67%, where linear SVM and random forest attained an accuracy of 60% and 46%. When considering demographic features alone, the linear SVM has 60% accuracy, our method has an accuracy of 60%, and random forest reaches 53% prediction accuracy.

II. Methods

A. Patients cohort

We obtained TCGA-GBM data for all white individuals that have tumor sample's binary alignment map (BAM) files, survival information (days from cancer index to death), and demographic features (age and gender) from NIH's GDC portal. A total of 272 patients met the inclusion criteria. We converted age, and gender into numerical values, and we also created an age group binary feature with 70 years threshold since GBM patients with age ≥ 70 have significantly lower survival time [17]. Table I shows GBM patients characteristics.

TABLE I
Cohort characteristics

	n=272
Short-/long-term survival	128/144
Average age	61.14 (±12.83)
Age ≥ 70	71
Male/female	177/95

B. SNPs calling and quality control

We performed variants calling from tumor samples only, and we used GATK HaplotypeCaller (version 3.8) [16]. GATK scans samples' genomes to identify regions with variability that exceed a defined threshold. From these regions, it builds an assembly directed graph with a reference genome as a template. It uses the most likely graph paths, the ones that have higher read data, to list candidate haplotypes. The candidate haplotype sequences are aligned against the reference genome with the Smith-Waterman algorithm to produce a CIGAR string. GATK determines the likelihood of haplotype by aligning every read against each haplotype with the PairHMM algorithm, which gives a likelihood for each haplotype given read data. From read data likelihoods, the program assigns allele likelihoods (possible genotypes). Finally, GATK uses Bayes' Theorem to assign genotypes for each sample from the list of possible genotypes.

We pooled all the samples together for variant discovery. To speed up the variants calling pipeline, we cut each chromosome into roughly 10 equal chunks and ran it at the same time on a cluster in a scatter-gather approach. In the final variant call set, we applied the GATK variant quality score recalibration (VQSR) algorithm, which uses machine learning to filter out low-quality variants. After applying VQSR filtering (soft filtering). We set the truth sensitivity filter for VQSR to a ”99.0%” threshold. We used the following annotations with VQSR to build a recalibration model: InbreedingCoeff, QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR. We also filtered out variants that have a depth (number of supporting reads) ≤ 5 or genotyping quality ≤ 20. We also removed non-SNPs variants and sites that have any missing value. The final output contains a matrix of SNPs and samples. Each SNP column is in the form A/B where A and B are the two alleles copies. We show the number of SNPs after applying each filtering method in table II.

TABLE II
TCGA-GBM SNPs count after applying three filtering methods

Filtering method	Number of SNPs
Soft filtering (VQSR)	304302
Hard filtering	155673
Soft+hard filtering	107777

C. SNPs encoding

To encode an SNP into a numerical format to perform machine learning tasks, we used the formula: $4 \times A + B$, where A and B are the two alleles copies for a given individual sample. We multiply A by 4 to consider all permutations in a multiallelic site (the maximum alternate alleles for an SNP is 3). For example, if an individual is homozygous at alternate allele 3 for a particular SNP, then this specific SNP encoding is 15. We sorted SNPs in increasing order according to their genomic position.

D. Training and test sets

To ensure the validity of results, we created separate training and test data sets. In the original TCGA-GBM dataset of 272 samples, we shuffled the data and randomly selected 5% from each class, to get a balanced subset, and kept this 10% balanced dataset for model testing. We used the remaining 90% for hyperparameters tuning, by employing 10-fold cross-validation, and to fit a model to predict the unseen test dataset with the best performing hyperparameters. Table III displays patients' characteristics in training and test data sets.

TABLE III
Training and test sets characteristics

	Training set n=244	Test set n=28
Survival $<$ 1 year	114	14
Survival $\geq$ 1 year	130	14
Average age	61.1 (±12.5)	61.4 (±14.6)
Age ≥ 70	62	9
Male/female	157/87	20/8

E. Hyperparameter selection

Classifiers hyperparameters, such as the SVM C regularization value, need to be set before model training begins, and thus are not optimized during the learning stage. To choose the best learning rate and the number of epochs hyperparameters for our neural network, we evaluated all possible pairs in the Cartesian product of the two sets: learning rate = $\{0.001, 0.01, 1\}$ and the number of epochs = $\{1,2,3, \ldots, 20\}$ using 10-fold cross-validation in the 90% of the original dataset (number of samples= 244) that we kept for training. We also employed the same method, with the same data in each fold, to select the best regularization C hyperparameter from the set C= {0.01, 0.1, 1} for linear SVM, as well as the number of trees to grow for random forest from the set {10, 100, 1000}. We then fit a model on the whole training dataset with the best performing hyperparameters and used the model built to predict the unseen 10% of the original data that we reserved as a test dataset.

F. Classifiers

1) Convolutional neural network: Convolutional neural networks (CNN) typically are stacked layers of convolution operations with pooling (downsampling of original data for training efficiency) and batch normalization layers in between convolutional layers. The convolution runs on sliding windows of a specified size and fixed step size, to control the moving dot product over training data. A non-linear and differentiable activation function, such as a rectified linear unit (relu), is then applied to the flattened output.

2) Multi-path model: Here we propose a new neural network system, where we feed the network two inputs: 1) SNPs data, 2) demographic data (age, age groups, and gender). Since we sorted SNP data in an increasing order based on its genomic position, we pass the SNPs through a series of 1D convolution, with different kernel sizes and a step size of 1, relu activation function, 1D average pooling, batch normalization layers. Simultaneously, we fed the three demographic features into two hidden-layers and merge the two paths and train the weights together through 3 fully connected layers. Then the network passes the weights into a sigmoid function that outputs a value between 0 and 1. If the output is ≥ 0.5, we assign it to class 1, and class 0 otherwise. We trained our model with stochastic gradient descent with a momentum that we set to 0.9. We used 10-fold cross-validation to select the number of epochs and learning rate (lr) value. We set the batch size to 128. Fig. 1 shows our multi-path model's architecture, all input and output shapes, and convolutions kernel and average pooling sizes. We implemented our model in Keras library [18]

3) Single-path model: We compare fitting a combined SNP and demographic features with our multi-path model to fitting a single-path 1D convolutional neural net with SNPs only and with three demographic features alone neural network.

4) Support vector machine: We used SVM with a linear kernel. Briefly, SVM finds a hyperplane that maximizes the distance between the two classes' data points that are closest to the margin (support vectors). In its soft-margin version, SVM allows misclassification of noisy data points and introduces a trade-off hyperparameter C that needs to be tuned. As C approaches infinity, the classifier gets closer to the hard-margin solution. We used 10-fold cross-validation to select the best performing C in the training dataset. We compared combining SNP and demographic features to fitting an SVM model with each data source alone. For SVM and random forest experiments, we used scikit-learn library [19]

5) Random forest: Random forest is an ensemble method that constructs many decision trees by choosing random samples with replacement to build each tree and randomly generates a subset of features to select from for each candidate split, usually the one with the highest Gini impurity or entropy, then it takes the majority vote of all trees predictions to output a class prediction. We used the default parameters for the quality measure of the split, and 10-fold cross-validation for the number of trees to construct. We fit a model with combined SNP and demographic features, SNP alone, and age+age group+gender alone.

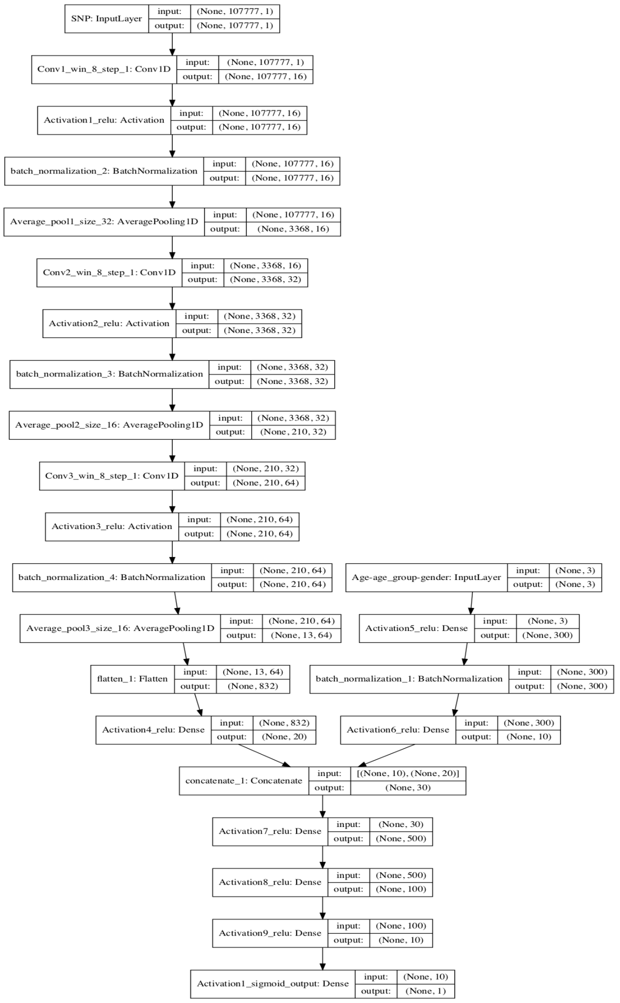

Fig. 1. Our proposed multi-path model architecture with SNP and demographic features inputs

G. Evaluation metrics

We used accuracy, which is the number of correctly classified samples over the number of all predicted samples, to measure classifiers' prediction power in the test data set. However, in training and validation data sets, we used the balanced accuracy, which is the average of true positive rate and true negative rate, since it has imbalanced class distribution.

III. RESULTS

A. Cross-validation

To tune classifiers hyperparameters, in training set we performed 10-fold cross-validation to select the best number of epochs and learning rates for single- and multi-path neural network system. Fig. 2 shows the mean balanced accuracy attained with different learning rates and the number of epochs across the ten folds. The best mean balanced accuracy of 63% (±0.08) across ten folds is realized when we fed both SNP and demographic features into our multi-path model with 0.01 as the learning rate. The mean balanced accuracy slightly drops after it reaches its peak at 13th epoch. With SNP data alone, the best learning rate was 0.001 with nine epochs, where the single-path convolutional neural network attained 54% (±0.12) mean balanced accuracy. With the demographic features alone, the single-path neural network reached its highest mean balanced accuracy of 59% (±0.12) at epoch 14 with a learning rate of 0.1.

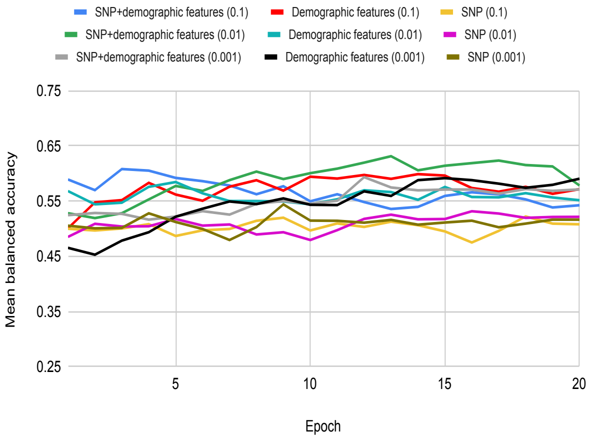

Fig. 2. Validation dataset average balanced accuracy across 10-folds as a function of the number of epoch and learning rate for multiple data inputs: demographic characteristics (age+age groups+gender) only, SNPs only, or SNPs and demographic characteristics combined. Each line color, which is shown in the series color legends, represents input data (learning rate in parentheses).

We also tuned SVM C regularization hyperparameter and the number of trees to grow for random forest classifiers. Fig. 3 shows that the SVM achieved its best results when C= $\{1, 0.1\}$ were both values are equally the best in combined SNP and demographic features, SNP alone, and demographic features alone. When learning with demographic features alone, SVM attained 61% (±0.08) mean balanced accuracy. SVM achieved 56% (±0.11) mean balanced accuracy with SNPs data alone, and the mean balanced accuracy drops to 50% (±0.10) when combining SNP and demographic features.

For random forest, setting the number of trees to 10 yielded a better performance for SNPs alone with 50% (±0.12) mean balanced accuracy and demographic features alone 52% (±0.08) mean balanced accuracy. The optimal number of trees for combined SNPs and demographic is 100 with 49% (±0.11) mean balanced accuracy. Fig. 3 shows the average 10-fold cross-validation with different hyperparameters for SVM and random forest.

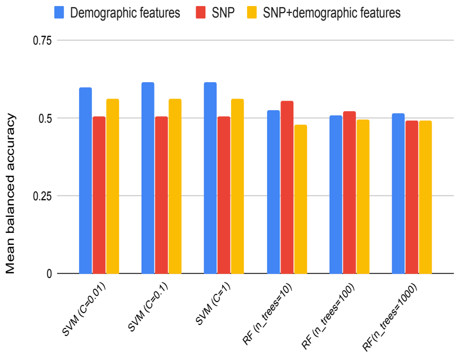

Fig. 3. Validation dataset mean balanced accuracy across 10-folds with linear SVM (with different C regularization values) and random forest (with different number of trees values) and multiple data inputs: demographic characteristics (age+age groups+gender) only, SNPs only, or SNPs and demographic characteristics combined. Each bar color represents a data source.

B. Test set prediction performance

After cross-validating the optimal hyperparameters for each classifier with each data source. We fit a model on the full training and validation sets and predict an independent, balanced test set. Table IV shows the accuracies attained by our model, SVM, and random forest accuracies with and without combining SNP and demographic features. Our multi-path model, with combined SNP and demographic features (age, age group, and gender), achieved the highest classification accuracy of 67%, when learning with the hyperparameters selected with the 10-fold cross-validation: learning rate of 0.01, and 13 epochs.

TABLE IV
PREDICTION ACCURACY ON TEST DATASET WITH THE OPTIMAL HYPERPARAMETERS

	SNP and demographic	SNP	Demographic
Our method	**0.67** lr =0.01 $epoch$ =13	**0.60** lr =0.001 $epoch$ =9	**0.60** lr =0.1 $epoch$ =14
SVM	0.60 $C = 1$	0.57) $C = 1$	**0.60** $C = 1$
Random forest	0.46 $ntrees$ =10	0.50 $ntrees$ =10	0.53 $ntrees$ =100

1) Combined SNP and demographic features: When combining SNP and demographic features, our multi-path model achieved an accuracy of 67%, which outperform both SVM (60%) and random forest (47%) accuracies. Furthermore, passing SNP, age, age groups, and gender yielded a nicer training curve that is stable across training epochs. Fig. 4 compares the training balanced accuracy of the combined SNP and demographic features with SNP data alone and demographics alone.

2) SNP and demographic features alone: In the test set, fitting a model with SNPs alone or age+age groups+gender alone had lower accuracy than combining SNPs and demographic features. With SNP data only, our single-path CNN

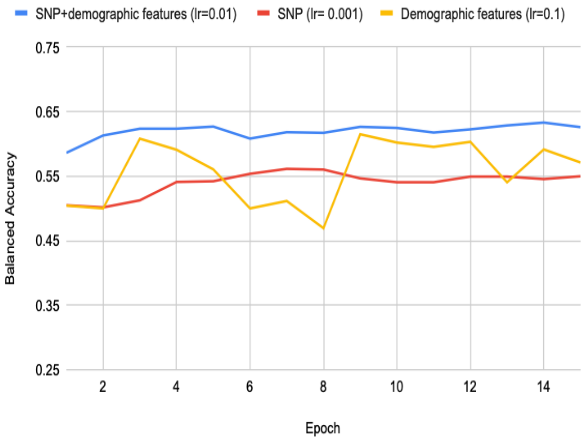

Fig. 4. Training accuracy on with training set (n=244)

had an accuracy of 60% with a learning rate of 0.001 and 9 epochs. SVM achieved an accuracy of 57% with C=1, and random forest accuracy is 50% with 100 trees. Fig. 5 displays our proposed model prediction accuracy with different data sources on the test set. With demographic features only, SVM and our single-path neural network performed equally with 60% accuracy. Random forest attained 50% accuracy. Table IV compares the accuracy achieved by our proposed CNN, SVM, and random forest with combined SNP and demographic features and with each data source alone.

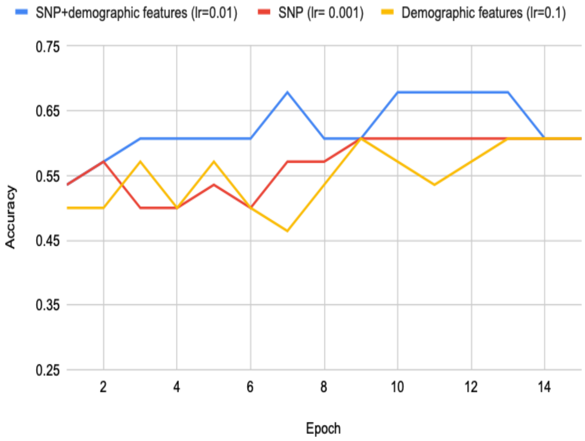

Fig. 5. Test set prediction accuracy for SNP and demographic features

IV. Conclusions

We proposed a new multi-path convolutional neural network for combined SNP and age, age group, and gender that improved upon SVM and random forest in terms of model accuracy in cross-validation and an independent test set. We show that using combined SNP and demographic features in a multi-path network attains a better classification performance than each data source alone and stabilized the learning process.

References

[1] P. Kao, T. Ngo, A. Zhang, J. W. Chen, and B. S. Manjunath, "Brain tumor segmentation and tractographic feature extraction from structural mr images for overall survival prediction," *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 128–141, 2019.

[2] W.-Z. Gao, L.-M. Guo, T.-Q. Xu, Y.-H. Yin, and F. Jia, "Identification of a multidimensional transcriptome signature for survival prediction of postoperative glioblastoma multiforme patients," *Journal of Translational Medicine*, vol. 16, no. 1, 2018.

[3] B. Liu, J. Liu, K. Liu, H. Huang, Y. Li, X. Hu, K. Wang, H. Cao, and Q. Cheng, "A prognostic signature of five pseudogenes for predicting lower-grade gliomas," *Biomedicine & Pharmacotherapy*, vol. 117, p. 109116, 2019.

[4] L. Macyszyn, H. Akbari, J. M. Pisapia, X. Da, M. Attiah, V. Pigrish, Y. Bi, S. Pal, R. V. Davuluri, and L. e. a. Roccograndi, "Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques," *Neuro-Oncology*, vol. 18, no. 3, pp. 417–425, 2015.

[5] Y. Tan, W. Mu, X.-c. Wang, G.-q. Yang, R. J. Gillies, and H. Zhang, "Improving survival prediction of high-grade glioma via machine learning techniques based on mri radiomic, genetic and clinical risk factors," *European Journal of Radiology*, 2019.

[6] A. Toft, T. Urup, I. J. Christensen, S. R. Michaelsen, B. S. Lukram, K. Grunnet, M. Kosteljanetz, V. A. Larsen, U. Lassen, and H. e. a. Broholm, "Abstract b3: Prognostic and predictive biomarkers in recurrent who grade 3 malignant glioma patients treated with bevacizumab and irinotecan," *Biomarkers*, 2015.

[7] G. Steponaitis, D. Skiriute, A. Kazlauskas, I. Golubickaite, R. Stakaitis, A. Tamaauskas, and P. Vaitkiene, "High chi3l1 expression is associated with glioma patient survival," *Diagnostic Pathology*, vol. 11, no. 1, 2016.

[8] R.-C. Chai, N. Wang, Y.-Z. Chang, K.-N. Zhang, J.-J. Li, J.-J. Niu, F. Wu, Y.-Q. Liu, and Y.-Z. Wang, "Systematically profiling the expression of eif3 subunits in glioma reveals the expression of eif3i has prognostic value in idh-mutant lower grade glioma," *Cancer Cell International*, vol. 19, no. 1, 2019.

[9] M. Tian, W. Ma, Y. Chen, Y. Yu, D. Zhu, J. Shi, and Y. Zhang, "Impact of gender on the survival of patients with glioblastoma," *Bioscience Reports*, vol. 38, no. 6, p. BSR20180752, 2018.

[10] A. Fogli, E. Chautard, C. Vaurs-Barrire, B. Pereira, M. Mller-Barthlmy, F. Court, J. Biau, A. A. Pinto, J.-L. Kmny, and T. e. a. Khalil, "The tumoral a genotype of the mgmt rs34180180 single-nucleotide polymorphism in aggressive gliomas is associated with shorter patients survival," *Carcinogenesis*, vol. 37, no. 2, pp. 169–176, 2015.

[11] A. Bunevicius, E. R. Laws, A. Saudargiene, A. Tamasauskas, G. Iervasi, V. Deltuva, T. R. Smith, and R. Bunevicius, "Common genetic variations of deiodinase genes and prognosis of brain tumor patients," *Endocrine*, 2019.

[12] D. Cui, J. Ren, J. Shi, L. Feng, K. Wang, T. Zeng, Y. Jin, and L. Gao, "R132h mutation in idh1 gene reduces proliferation, cell survival and invasion of human glioma by downregulating wnt/-catenin signaling," *The International Journal of Biochemistry & Cell Biology*, vol. 73, pp. 72–81, 2016.

[13] K. Wang, Y. Wang, X. Fan, J. Wang, G. Li, J. Ma, J. Ma, T. Jiang, and J. Dai, "Radiological features combined with idh1 status for predicting the survival outcome of glioblastoma patients," *Neuro-Oncology*, vol. 18, no. 4, pp. 589–597, 2015.

[14] H. Yan, D. W. Parsons, G. Jin, R. McLendon, B. A. Rasheed, W. Yuan, I. Kos, I. Batinic-Haberle, S. Jones, and G. J. e. a. Riggins, "Idh1 and idh2 mutations in gliomas," *New England Journal of Medicine*, vol. 360, no. 8, pp. 765–773, 2009.

[15] S. E. Combs, S. Rieken, W. Wick, A. Abdollahi, A. von Deimling, J. Debus, and C. Hartmann, "Prognostic significance of idh-1 and mgmt in patients with glioblastoma: One step forward, and one step back?" *Radiation Oncology*, vol. 6, no. 1, 2011.

[16] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, and M. e. a. Daly, "The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[17] U. Smrdel, M. S. Vidmar, and A. Smrdel, "Glioblastoma in patients over 70 years of age," *Radiology and Oncology*, vol. 52, no. 2, pp. 167–172, 2018.

[18] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.