frontiers in Genetics

ORIGINAL RESEARCH
published: 28 March 2022
doi: 10.3389/fgene.2022.855420

A Deep Learning–Based Framework for Supporting Clinical Diagnosis of Glioblastoma Subtypes

Sana Munquad[1], *Tapas Si*[2], *Saurav Mallik*[3], *Asim Bikas Das*[1]* *and Zhongming Zhao*[3,4,5]*

[1]Department of Biotechnology, National Institute of Technology Warangal, Warangal, India, [2]Department of Computer Science and Engineering, Bankura Unnayani Institute of Engineering, Bankura, India, [3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States, [4]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States, [5]Department of Pathology and Laboratory Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, United States

Understanding molecular features that facilitate aggressive phenotypes in glioblastoma multiforme (GBM) remains a major clinical challenge. Accurate diagnosis of GBM subtypes, namely classical, proneural, and mesenchymal, and identification of specific molecular features are crucial for clinicians for systematic treatment. We develop a biologically interpretable and highly efficient deep learning framework based on a convolutional neural network for subtype identification. The classifiers were generated from high-throughput data of different molecular levels, i.e., transcriptome and methylome. Furthermore, an integrated subsystem of transcriptome and methylome data was also used to build the biologically relevant model. Our results show that deep learning model outperforms the traditional machine learning algorithms. Furthermore, to evaluate the biological and clinical applicability of the classification, we performed weighted gene correlation network analysis, gene set enrichment, and survival analysis of the feature genes. We identified the genotype–phenotype relationship of GBM subtypes and the subtype-specific predictive biomarkers for potential diagnosis and treatment.

Keywords: deep learning, glioblastoma multiforme, biomarkers, co-expression gene module, machine learning

OPEN ACCESS

Edited by:
Zheng Xia,
Oregon Health and Science University,
United States

Reviewed by:
Xiangnan Guan,
Genentech, Inc., United States
Fuhai Li,
Washington University in St. Louis,
United States

Correspondence:
Asim Bikas Das
asimbikas@nitw.ac.in
Zhongming Zhao
zhongming.zhao@uth.tmc.edu

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *15 January 2022*
Accepted: *17 February 2022*
Published: *28 March 2022*

Citation:
Munquad S, Si T, Mallik S, Das AB and Zhao Z (2022) A Deep Learning–Based Framework for Supporting Clinical Diagnosis of Glioblastoma Subtypes. Front. Genet. 13:855420. doi: 10.3389/fgene.2022.855420

INTRODUCTION

Glioblastoma multiforme (GBM), which is the grade IV of glioma as defined by the World Health Organization (WHO), is a highly invasive and devastating primary form of brain cancer. The complexity and molecular heterogeneity of GBM pose the challenge for accurate diagnosis and therapy (Verhaak et al., 2010; Zhang et al., 2020). The prognosis for patients with GBM is poor, and median survival is 12 months (Witthayanuwat et al., 2018). Because of enormous molecular heterogeneity and difficulty in early diagnosis, the molecular mechanisms of GBM tumorigenesis are not clear. This leads to ineffective therapeutic intervention, and many patients relapse. However, with the current treatment options, i.e., surgery, radiotherapy, and chemotherapy, patient life expectancy can be increased, but these are not curative. To find the remedial solution, understanding the molecular features and identification of GBM subtypes is crucial. An earlier study shows that GBM can be classified into four subtypes based on transcriptional features, i.e., classical, neural, proneural, and mesenchymal. However, recent findings suggest that the neural subtype probably arises due to the contamination of normal neuronal tissue tumor margins (Wang et al., 2018).

Therefore, GBM is currently classified into three subtypes. There are many other studies to find other subtypes using omics and clinical data (Park et al., 2019). Histopathological-based diagnosis is the most common method for subtype identification. However, it often leads to the inaccurate classification of subtypes due to interobserver variability (Van den Bent, 2010). Accurate pathological subtype diagnosis is pivotal for optimal patient management. Because GBM subtypes are histologically and genetically heterogeneous, they differ in gene expression, mutation, and epigenetic states, which lead to different therapeutic responses and clinical outcomes (Brennan et al., 2013; Zhang et al., 2020).

Recent advances of sequencing technologies have helped generate massive omics data in cancer, leading to a deep understanding of the molecular mechanisms in both common and rare cancers (Mardis and Wilson, 2009; Campbell et al., 2020). Data from sequencing experiments reveal that cancer initiation, progression, and maintenance are caused by the perturbations in multiple genomics and epigenomics factors. Additionally, genomics and epigenomics biomarkers have emerged as promising tools for developing the precision medicine and stratification of cancer subtypes and grades (Aran et al., 2013; Jurmeister et al., 2019; Jayanthi et al., 2020; Yoon et al., 2021). Alteration of gene expression and DNA methylation is the most prominent genomic and epigenomic event in cancer cells (Chakravarthi et al., 2016). The genome-wide analysis reveals that changes in gene expression and methylation patterns in several positions in the genome are strongly associated with GBM formation and progression (Bozdag et al., 2013; Dong and Cui, 2019; Vinel et al., 2021). Gene expression and methylation are both strongly interlinked processes; methylation levels in promoter regions influence the gene expression by regulating transcription factor binding (Mallik et al., 2020b). On many occasions, hypermethylation of CpG sites on promoter regions inhibits gene expression, whereas hypomethylation causes higher expression of genes (Moore et al., 2012). Therefore, classification using multiple "omics" data, i.e., transcriptome and methylome, can provide optimal features for the clinical diagnosis of cancer subtypes. However, enormous amounts of genetic and epigenetic alterations pose challenges to finding the unique molecular marker for diagnosing GBM subtypes. Benefitting from recent advances in computational methods, such as deep learning (DL) and traditional machine learning (ML), it is possible to scan the genome-wide transcriptome and methylome data to find the subtype-specific molecular feature for diagnosis (Qin et al., 2020).

We have implemented ML and DL algorithms for the precise and accurate classification of GBM subtypes in the present work. Each data type (i.e., transcriptome and methylome) and its integrated subsystem were separately used for classification. We found that the performance of the convolutional neural network (CNN) was superior (always >90%) compared with the other ML models. In addition, we examined the biological relevancy of features using weighted gene co-expression network analysis (WGCNA) and Gene Ontology (GO) analysis. Results show distinct co-expression modules are linked to each GBM subtype and are associated with subtype-specific biological functions. Moreover, several genes in the co-expression module are associated with patients' survival. Overall, our findings suggest that a combination of LASSO feature selection and CNN can classify the subtype of GBM with higher accuracy and be used for clinical diagnosis.

MATERIALS AND METHODS

Data Collection, Preprocessing, and Integration

In this study, we analyze TCGA GBM transcriptome (RNA-seq) and methylome (Illumina Infinium HumanMethylation450 platform) data. The data set was retrieved from UCSC Xena (https://xena.ucsc.edu/) (Goldman et al., 2020). Log2 (RSEM +1) (RSEM: RNA-Seq by Expectation Maximization) values for transcriptome, and β values for methylation were used for analysis. Next, the low-expression genes were removed from transcriptome data [log2 (RSEM +1) <0.1 in 90% sample], and data was scaled before analysis. Based on the clinical information, patients (n = 155) were divided into three categories based on cancer subtype, i.e., classical (n = 42), mesenchymal (n = 55), and proneural (n = 39) for transcriptome data. Similarly, based on the clinical information, we divided the methylome data (n = 84) into a particular subtype, i.e., classical (n = 29), mesenchymal (n = 32), and proneural (n = 23). Next, based on the clinical information, patients with both transcriptome and methylome profiles in TCGA were screened to integrate the transcriptome and methylome data. The total number of these patients with omics data was 52, including classical (n = 16), mesenchymal (n = 22), and proneural (n = 14). Due to the unavailability of healthy patient data for both transcriptome and methylome, we used the Z-score to classify higher and lower expression of genes and hyper- and hypo-methylated CpG sites. We calculated the Z-score for each gene or CpG site in a specific subtype using the following formula:

$$Z - score = \frac{\bar{x} - \mu}{\sigma}.$$

Here, $\bar{x}$ represents subtype-specific average expression or methylation level of a gene/CpG site, and μ and σ represent the population mean and population standard deviation, respectively (Bandyopadhyay et al., 2014). We applied Z-score>1 for higher expression and hypermethylation and Z-score < -1 for lower expression and hypomethylated on each subtype of GBM. Next, we screened the higher and lower expressed genes whose promoter regions were differentially methylated, considering that the differential methylation in the promoter regions may alter the corresponding gene's expression. Finally, genes with both differential expression patterns and differential methylation promoter regions were used for further analysis (Maegawa et al., 2010; Sumithra et al., 2019). We collected the external data set from the Gene Expression Omnibus (GEO) repository for validation. GSE145645 was used to validate the model constructed using transcriptome and integrated data. GSE145645 contained all the subtypes of

GBM, i.e., classical (n = 9), mesenchymal (n = 14) and proneural (n = 9). Models built on methylome data were further validated using GSE128654, which consisted of classical (n = 11), mesenchymal (n = 8), and proneural (n = 10) subtypes.

Clustering Using t-SNE and Principal Component Analysis

The subtype-specific clustering of patients using transcriptome, methylome, and integrated data was visualized by t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA) (Van der Maaten and Hinton, 2008). t-SNE was performed using the TSNE package in Python. For each t-SNE run, 1000 embeddings were created. Apart from that, we used PCA for better visualization of GBM subtypes; the ggfortify and cluster packages in R were used.

Features Selection

We performed feature or variable selection to improve the performance of ML and DL algorithms. The least absolute shrinkage and selection operator (LASSO) was performed on all types of preprocessed data (Muthukrishnan and Rohini, 2016). We used default parameter values for lambda (the tuning factor that controls the strength of penalty) and dropped those genes having the coefficient value of zero. LASSO was implemented in the ScikitLearn (https://scikit-learn.org) package in Python.

ML and DL Models for Classification of GBM Subtypes

We performed classification on the subtype of GBM as a multiclassification problem using gene expression levels as covariates. Several ML and DL algorithms were used for classification: support vector machine (SVM), random forest (RF), naïve Bayes (NB), logistic regression (LR), k-nearest neighbors (kNN), and CNN. SVM is used for the classification between the classes to find the optimal hyperplane (Afifi et al., 2017). The optimal hyperplane boundary not only separates the classes, but also maximizes the margin between the classes. The margin is the longest distance between the hyperplane and the nearest data (support vector) in each class. RF is a tree-based ensemble learning method that constructs several decision trees and gives the output for classification based on a majority vote between the estimators (trees). Gaussian NB classifier is an easy and simple Gaussian distribution that is dependent on the application of the Bayes theorem (Kaviarasi and Gandhi Raj, 2019). In Gaussian NB, each variable is considered as an independent variable and trained efficiently in supervised learning. It requires small measures of training data, which are essential for characterization and necessary for classification. A logistic regression classifier predicts the response based on one or more predictor variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. kNN (Liu et al., 2012) is a clustering algorithm that is widely used for pattern classification based on similarity measures. It utilizes standard Euclidean distance and evaluates the distinguishing features. kNN estimates the class attribute depending upon a neighborhood of close (or similar) patterns in the feature space. CNN is one of the deep feed-forward artificial neural network architectures that consist of the convolutional layer, activation function, and pooling layer. Convolution is one type of linear operation used instead of general matrix multiplication in convolution layers where filters are applied to original data or to feature maps in deep CNN. The convolution operation (denoted by an asterisk) is defined by

$$f(t) = (x*K)(t),$$

where the function $x\ (t)$ is referred to as input, $K(t)$ is referred to as kernel, and $f(t)$ is referred to as output. In this paper, all ML classifiers on the Python platform use the sklearn library. The Keras library was used to construct the model architecture for CNN. Eight convolutional layers were used for obtaining the best result. All parameters for CNN are provided in **Supplementary Table S1**. Furthermore, parameters were optimized by the grid search method using the GridSearchCV package in Python. After obtaining optimal features, stratified k-fold was applied on the 70% training data set, and average performance measures were recorded. In stratified k-fold CV, the data set is divided into k independent folds, where k-1 folds were used to train the network, and the remaining one is reserved for test purposes. This procedure is then repeated until all folds are used once as a test set. The final output is then computed by averaging over the obtained performance parameters from each test set.

Performance Evaluation

The performance of ML and deep learning models was evaluated using accuracy, recall, precision, F1-score, FPR, GM, and MCC. At first, we generated a confusion matrix to compute these performance scores. The confusion matrix is a table that categorizes the model's prediction of whether it matches the actual value. We calculated true positive (TP), true negative (TN), false positive (FP), false negative (FN) from the confusion matrix (Mallik and Zhao, 2020). Then, we calculated the accuracy or success rate as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

The sensitivity or TP rate of an ML model was measured using the following equation:

$$Sensitivity = \frac{TP}{TP + FN}.$$

The specificity or TN rate of an ML model was measured using the following equation:

$$Specificity = \frac{TN}{TN + FP}.$$

The precision or positive predicted value was measured using the following equation:

$$Precision = \frac{TP}{TP + FP}.$$

A measure of model performance that combines precision and recall into a single number is known as the F measure or F1-score. The following equation was used to compute the F1-score:

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN}.$$

Geometric mean (GM) is the average value or mean, which signifies the central tendency of the set of numbers by taking the *n*th root of the product of their values.

$$Geometric\,mean = (x_1, x_2 \ldots\ldots x_n)\frac{1}{n}.$$

Mattews correlation coefficient (MCC) measures the correlation of the true classes with the predicted labels.

$$MCC = \frac{(TP*TN) - (FP*FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

We used the sklearn metrics library in Python to calculate the above score by importing functions such as confusion_matrix and classification_performance. Finally, we visualized the model performance across a wide range of conditions using receiver operating characteristic curve (ROC) plots using the roc_curve function.

Ranking of the Model

Algorithm performance was compared using multi-criteria decision analysis (MCDA)/multi-criteria decision making (MCDM). The technique for order of preference by similarity to ideal solution (TOPSIS), an established MCDM method, was used to rank. Multiple criteria, such as accuracy, sensitivity, precision, G-mean, F-measure, FPR, and MCC, were used in TOPSIS (Triantaphyllou, 2000).

Weighted Correlation Network Analysis

We identified co-expressed gene modules and analyzed the module-trait relationship using the WGCNA package in R (Langfelder and Horvath, 2008). First, the similarity matrix between each pair of feature genes in a specific subtype was measured based on Pearson's correlation coefficient. Next, we transformed the similarity matrix into an adjacency matrix. The soft power β value was calculated for building the proximity matrix so that the co-expression network conformed to a scale-free network based on connectivity. Subsequently, we computed the topological overlap matrix (TOM) and the corresponding dissimilarity (1-TOM) value. Next, a dynamic tree cut algorithm was implemented to detect gene co-expression modules. The co-expression modules were constructed with a cut height of 0.6, and a minimum module size was set to 10 (transcriptome), 10 (methylome), and 5 (integrated) genes, respectively.

Gene Set Enrichment and Survival Analysis

We performed the biological process and functional enrichment analysis using Enrichr (Kuleshov et al., 2016). Terms were considered statistically significantly enriched if the adjusted *p*-value was less than 0.05. The gene list from each positively correlated module was used to examine the enrichment of GO biological processes and molecular function terms. We performed overall survival and log-rank test of a co-expressed module using the survminer and survival package in R. We calculated the average expression of all genes in the module. Survival was compared between two groups: patients with higher (⩾75 percentile) and lower (⩽25 percentile) gene expression levels. Furthermore, we performed the overall survival analysis of specific genes using GEPIA (Tang et al., 2017). GEPIA performs survival analysis based on The Cancer Genome Atlas (TCGA) gene expression levels and patient clinical information. Here, the TCGA GBM data set was used for survival analysis. GEPIA generates Kaplan–Meier plots and performs the log-rank test to identify the genes associated with patient survival.

RESULTS

The etiology of GBM is associated with the alteration of transcriptome and methylome patterns. Therefore, the multi-omics approach that combines genome-wide methylation with transcriptome (RNA-seq) data can provide novel insights into biological function and disease mechanisms. In this work, we first separately analyzed the transcriptome and methylome, and then we integrated both data types to identify the molecular feature and classify the GBM subtypes.

Classification of GBM Subtype Using Transcriptome

The transcriptome data of the GBM at TCGA contained 20,531 genes. After removing the low-expression genes, a total of 14,125 genes were found expressed in all GBM subtypes, including classical (n = 42), mesenchymal (n = 55), and proneural (n = 39). These genes were taken for further analysis. However, 14,125 genes could not be used as variables for prediction as the data is high-dimensional, leading to the inaccurate classification of subtypes. Therefore, we performed the LASSO to reduce the dimension of data and subsequently for selecting top key feature genes to enhance the prediction accuracy of the DL and ML model. LASSO performs $L1$ regularization and adds a penalty to the loss function. This penalty contains the absolute value of the regression coefficients. It attempts to minimize the cost function and automatically selects relevant features that are useful, and the remaining features are discarded with a coefficient equal to zero. The coefficients of the regression variables having nonzero values were selected as an optimal feature for further processing. A total of 201 feature genes were obtained after performing the LASSO analysis (**Supplementary Table S2**). Next, we performed t-SNE and PCA to examine the local structure of data, including 14,125 genes and 201 feature genes. We observed improved subtype-specific separation between patients using 201 feature genes compared to 14,125 genes, indicating that the LASSO feature selection method efficiently extracted most variable features from the transcriptome data (**Figures 1B–E**). Additionally, the percentage of variability in principal component 1 (PC1) was increased in the PCA of 201 feature genes compared with the preprocessed data (**Figures 1D,E**). These results indicate that

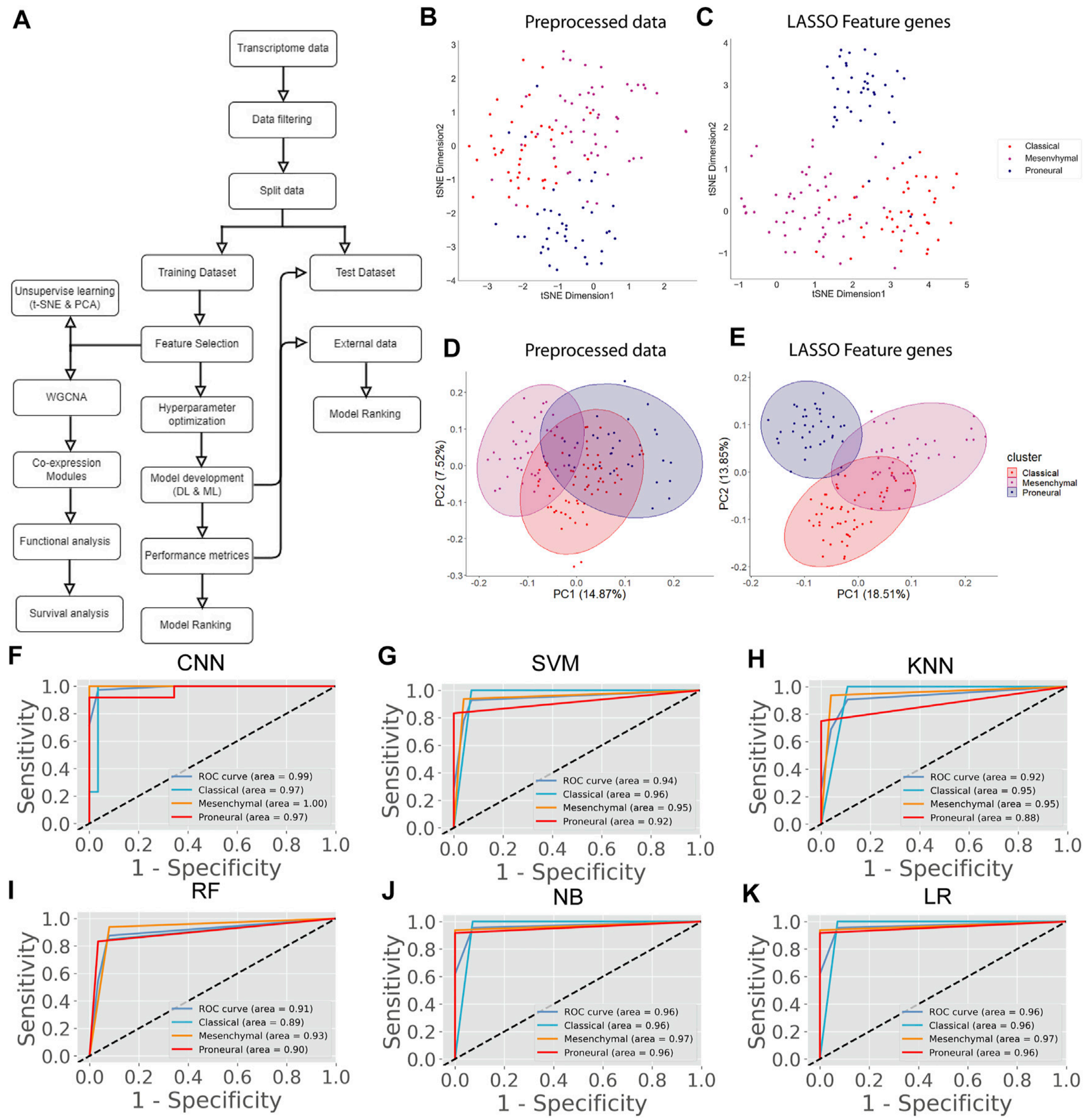

FIGURE 1 | GBM subtype classification using transcriptome data. **(A)** The flow chart shows DL and ML pipelines using genome-wide transcriptome data to classify the subtypes. **(B, C)** t-SNE plots to visualize the subtype-specific clustering of the patients using preprocessed data and features genes. **(D, E)** PCA plots to visualize the subtype-specific clustering of patients using preprocessed data and feature genes. **(F–K)** ROC of various prediction models. ROC plots were generated using a test data set.

information contained in 201 feature genes could separate the subtype with higher accuracy upon implementing DL and ML algorithms. However, distinct clusters of subtypes were not formed either in t-SNE or PCA.

Next, we proceeded to apply DL (CNN) and ML algorithms (i.e., SVM, KNN, RF, NB, LR) to classify subtypes of GBM using these feature genes as variables. We divided the data into training (70%) and test (30%) data sets. Seventy percent of the data was used for parameter optimization and to assess the performance of each model. The remaining 30% of data was used for independent predictors. Additionally, an external data set was also used for the final validation of models (**Figure 1A**). In the model training step, 70% of the data was used to obtain the best combination of hyperparameters using the grid search method for each DL and ML model. Next, we performed the stratified k-fold cross-validation (k = 10) on the training data using the optimal hyperparameters obtained from the grid search and recorded average performance measures of each model (**Table 1**). The performance of the models was evaluated using average accuracy, recall, precision, F1-score, FPR, GM, and MCC (see materials and methods). We observed that the prediction accuracy of CNN was superior (98.56%) compared with the other ML models. Even standard deviation (±0.03) and FPR (0.01) were minimum in the case of CNN. The MCC score is 0.97 for CNN, which represents

TABLE 1 | Models performance and ranking for transcriptome data.

Method	Performance measures (Average of tenfold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	91.42% (±0.08)	84.48	91.80	85.51	0.06	91.52	0.82	4
KNN	91.03% (±0.06)	85.78	90.59	86.06	0.07	91.44	0.82	5
RF	93.06% (±0.08)	88.52	93.04	89.15	0.05	93.02	0.85	3
NB	90.15% (±0.07)	86.08	87.16	85.38	0.08	90.52	0.80	6
LR	93.32% (±0.05)	89.47	92.12	89.97	0.05	93.61	0.86	2
CNN	98.56% (±0.03)	97.86	98.36	97.81	0.01	98.64	0.97	1

TABLE 2 | Models performance and ranking for validation data (transcriptome).

Method	Performance measures (Average of tenfold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	79.14% (±0.14)	71.33	63.57	65.68	0.11	84.07	0.71	4
KNN	79.15% (±0.14)	71.33	63.57	65.68	0.11	84.07	0.71	5
RF	80.57% (±0.22)	71.38	65.75	67.54	0.10	85.85	0.66	3
NB	77.59% (±0.17)	68.28	61.90	64.02	0.12	82.99	0.68	6
LR	81.20% (±0.15)	74.68	66.01	68.90	0.10	86.44	0.75	2
CNN	92.70% (±0.12)	90.20	88.77	89.24	0.01	98.25	0.96	1

the excellent correlation between the observed and predicted classifications. We observed that the performance of other ML classifiers was also good (accuracy >90%). Therefore, to compare the overall performance, we performed MCDM using TOPSIS (Si et al., 2021). All performance measures mentioned in **Table 1** were considered for the ranking, and CNN topped the overall ranking. To validate this observation, we performed the classification using two data sets, i.e., 30% data as the test data (or independent data) and an external data set from GEO (GSE145645). In the test data, the prediction accuracy (98.56%) of CNN was superior to other ML models and the MCC score was 0.96 (**Supplementary Table S3**). It is always desirable to have a highly sensitive and highly specific model for diagnosis. Therefore, we visualized the relationship between sensitivity and specificity using the ROC curve (**Figures 1F–K**). The ROC curve represents the probability of a TP result or the test's sensitivity against the probability of an FP result for a range of different cutoff points. **Figure 1F** shows the area under the ROC curve (AUC) is 0.99 for CNN, indicating that CNN can classify the GBM subtype with high specificity and sensitivity for clinical diagnosis. Additionally, classification with the external data set also represented a similar outcome; i.e., the performance of CNN was higher (**Table 2**). While validating with the external data set, we implemented tenfold cross-validation to calculate the average performance measure and compared the model performance by computing the rank. Furthermore, we compared the classification accuracy of the LASSO feature with the features selected using the variance. Gene with higher variance may contain more useful information. We selected the top 201 variable genes according to the degree of variance across all samples to compare the performance with LASSO. We performed the CNN using the same parameters and tenfold cross-validation. The average accuracy was 84.02% (±0.08). Therefore, the accuracy of prediction was less than LASSO features (98.56%). Hence, model building to validation, we observed that the feature genes from LASSO and CNN were the best for subtype classification for the transcriptome data. Therefore, we implemented this framework in subsequent analysis.

Classification of GBM Subtype Using Methylome

In the previous section, we classify the GBM subtype using the transcriptome data (or gene expression data) because the alteration of gene expression is a hallmark of oncogenesis. However, the level of gene expression is regulated by DNA methylation. Therefore, changes in DNA methylation patterns can play a crucial role in GBM development. Recent studies show that DNA methylation biomarkers are essential for improving and designing cancer therapy (Locke et al., 2019). Hence, the information contained in methylation data could possibly help to classify the GBM subtype. The genome-wide methylation or methylome data of 84 GBM patients were retrieved from the UCSC Xena database. We selected the data from the Illumina Infinium HumanMethylation450 platform (450K array) that has 4,85,577 probe sites. In this data set, the methylation level is estimated using the beta value. The beta value ranges from zero to one, representing the ratio of the intensity of the methylated bead type to the combined locus intensity. Thus, higher beta values represent a higher level of DNA methylation, i.e., hypermethylation and lower beta values represent a lower level of DNA methylation, i.e., hypomethylation. The recent reports show that the hypermethylation/hypomethylation level in the promoter region (e.g., defined as TSS1500 upstream to TSS200 downstream of TSS, 5′UTR, and first exon; TSS denotes

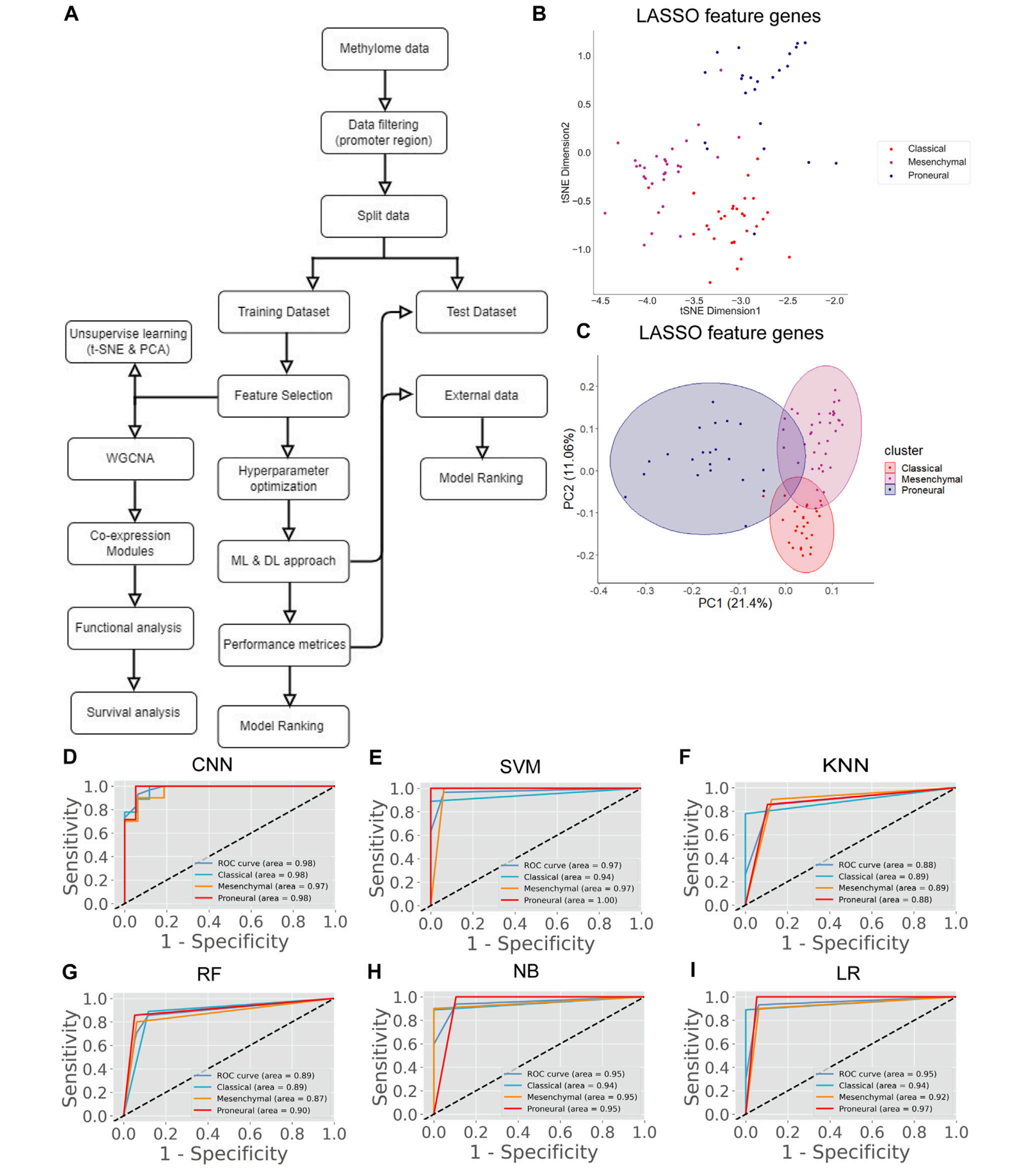

FIGURE 2 | GBM subtype classification using methylome data. **(A)** The flow chart shows DL and ML pipelines using genome-wide DNA methylation data to classify the subtypes. **(B, C)** t-SNE and PCA plots to visualize the subtype-specific clustering of the patients from features gene. **(D–I)** ROC of various prediction models. ROC plots were generated using the test data set.

transcription start site) and gene body determine the gene expression level (Sandoval et al., 2011; Yang et al., 2014). Therefore, we screened the promoter and gene body methylation data to perform classification because the alteration of methylation levels in these regions can influence the gene expression level and subsequently influence the

TABLE 3 | Models performance and ranking for methylation data.

Method	Performance measures (Average of tenfold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	90.61% (±0.09)	86.40	87.67	84.49	0.07	90.55	0.81	4
KNN	90.72% (±0.12)	85.86	88.10	84.90	0.07	90.36	0.81	5
RF	91.03% (±0.10)	86.92	89.74	86.33	0.06	90.81	0.82	3
NB	92.34% (±0.08)	88.85	92.63	88.46	0.05	92.03	0.84	2
LR	89.84% (±0.11)	83.71	82.70	81.80	0.08	89.46	0.78	6
CNN	97.54% (±0.05)	96.77	97.71	96.47	0.01	97.47	0.95	1

TABLE 4 | Models performance and ranking for external data (methylation).

Method	Performance measures (Average of tenfold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	82.42% (±0.23)	76.65	73.58	74.60	0.09	85.26	0.76	4
KNN	79.09% (±0.20)	68.00	63.19	64.22	0.13	81.49	0.66	6
RF	82.81% (±0.16)	76.27	70.31	72.29	0.08	88.19	0.78	3
NB	81.52% (±0.15)	71.46	65.50	66.91	0.11	83.45	0.71	5
LR	87.42% (±0.17)	84.34	81.08	82.17	0.05	92.92	0.86	2
CNN	91.91% (±0.13)	90.50	89.15	89.60	0.01	97.63	0.96	1

biological processes (Dhar et al., 2021). The CpG sites, which include all promoter regions and the gene body, were screened for feature selection. By using LASSO, we obtained 498 features CpG sites (**Supplementary Table S2**). Next, we examined the subtype-specific clustering of patients with these 498 features CpG sites using t-SNE and PCA. Results show that there was slighter mixing among the different subtypes (**Figure 2B,C**). Next, we performed the DL and ML using these 498 CpG sites as variables. We repeated the same methodology as described in the previous section. First, the methylome data were divided into training (70%) and test (30%). The hyperparameters were optimized using the grid search method, and tenfold cross-validation was performed on the training data. The average performance measures were used to select the top-performing model using MCDM (**Figure 2A**). The overall performance of CNN was superior compared with other ML models using methylation data as well (**Table 3**). Next, we validated our observation with the 30% test data set (**Supplementary Table S4**) and an external data set (GSE128654) (**Table 4**). ROC plots (**Figures 2D–I**) show that the performance of the CNN (AUC = 0.98) was better compared with other ML models. However, the accuracy value is 89.0%, which is lower than the ML models. The overall performance of CNN on external data is superior (Rank = 1, see **Table 4**). These results indicate that CNN is the best classifier for predicting the GBM subtype using DNA methylation data.

Classification of GBM Subtype by Integrating the Methylation and Transcriptome Data

There are several studies where only one type of "omics" data is used, such as either gene expression or methylation data, to identify the biomarkers or classify the cancers (Díaz-Uriarte and Alvarez de Andrés, 2006; Wang et al., 2020). However, DNA methylation and gene expression are integrated processes that determine cellular fate (Basu and Tiwari, 2021). The perturbation of gene expression in many human cancers is due to the change of methylation pattern (Langfelder and Horvath, 2008). Hence, integrating these strongly interlinked cellular processes and subsequent analysis could facilitate finding a more effective diagnostic option (Mallik et al., 2020a). The patients having both transcriptome and methylome data were selected for data integration. Next, we screened the gene and methylation sites based on *z*-score, i.e., $z > 1$ and $z < -1$ (see materials and methods). A *z*-score greater than 1 or less than -1 indicates the expression and methylation is greater or less than the population mean. We identified common genes whose expression and methylation both are $z > 1$ or $z < -1$ in each subtype. Next, we combined all these genes ($n = 4{,}231$) and used their gene expression level to find the most variable features ($n = 75$) using LASSO (**Supplementary Table S2**). We observed that 75 feature genes form the distinct subtype-specific clusters with PCA and t-SNE (**Figure 3B–E**). Compared with previous features from transcriptome and methylome data, the feature genes of the integrated data significantly improve the clustering of the GBM subtype. Next, we implemented CNN using these feature genes and compared CNN performance with the other five ML algorithms (**Figure 3A**). In this case, the CNN performance was also ranked on top (**Table 5**). Furthermore, we validated the model with 30% test data (**Supplementary Table S5**) and external data (**Table 6**). ROC plots generated using test data explain the decent performance of CNN (AUC = 0.91 and accuracy = 87.50%) (**Figures 3F–K**). The validation with external data showed that CNN was the top performer (accuracy = 94.48%) for classification (**Table 6**). It can be concluded that in all three types of analysis,

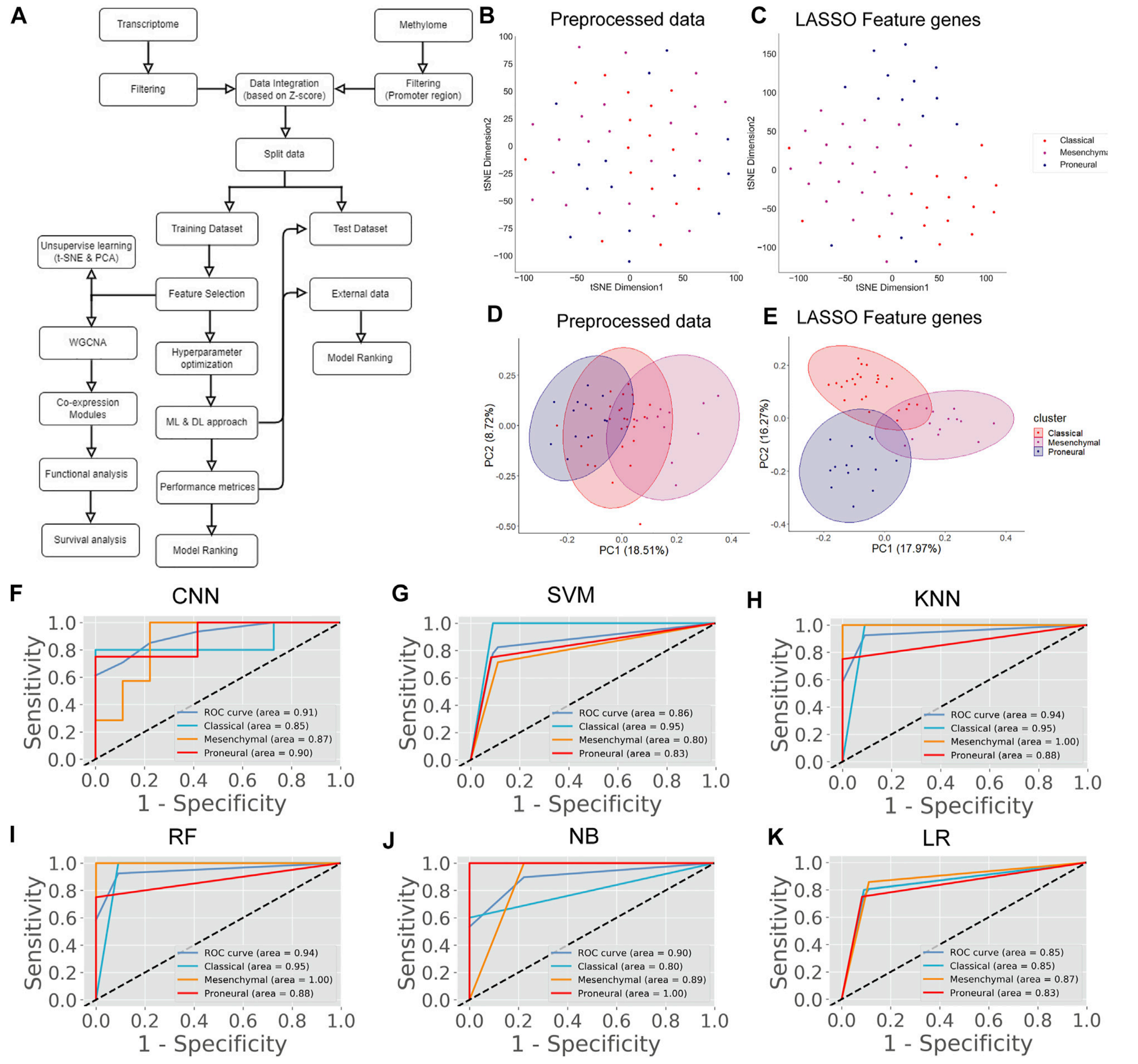

FIGURE 3 | GBM subtype classification using integrated data. **(A)** The flow chart shows DL and ML pipelines using the integrated data of transcriptome and methylome to classify the subtypes. **(B,C)** t-SNE plots to visualize the subtype-specific clustering of patient from features gene. **(D,E)** PCA plots to visualize the subtype-specific clustering of patient from features gene. **(F–K)** ROC of various prediction models. ROC plots were generated using the test data set.

CNN efficiently classified the GBM subtypes. However, the features from integrated data specifically cluster the subtype of GBM with PCA and t-SNE. Moreover, the consistent all-around performance of CNN proves that CNN can be used as a computational tool for the clinical diagnosis GBM subtype.

The Biological Relevance of Features and Identification of Biomarkers

In the preceding steps, we extracted features from large-scale transcriptome and methylome data sets to develop the predictive tool for subtype identification. We observed that selected features from each type of data have excellent separability power, and therefore, we achieved classification accuracy >90% in every case. This indicates that any subset of these features is probably associated with a particular subtype (or phenotype). Therefore, further analysis of these features genes can link the genotype to phenotype. We performed WGCNA to understand genotype-to-phenotype relationships. WGCNA can find the module of highly correlated genes and their association with a specific subtype of GBM (Langfelder and Horvath, 2008). We constructed the co-expression module using the feature genes expression from

TABLE 5 | Models performance and ranking for integrated data.

Method	Performance measures (Average of tenfold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	89.94% (±0.10)	86.47	81.11	81.65	0.07	90.02	0.82	5
KNN	91.87% (±0.13)	88.35	82.68	84.57	0.06	91.81	0.84	3
RF	93.67% (±0.10)	88.70	84.63	86.06	0.04	93.52	0.89	2
NB	89.95% (±0.14)	83.16	77.12	79.14	0.08	89.43	0.79	6
LR	92.18% (±0.10)	87.10	81.38	83.43	0.06	91.77	0.85	4
CNN	98.20% (±0.05)	98.44	97.97	97.60	0.01	98.25	0.97	1

TABLE 6 | Models performance and ranking for external data (transcriptome).

Method	Performance measures (Average of tenfold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	63.15% (±0.12)	46.43	35.70	37.89	0.22	68.38	0.38	6
KNN	67.08% (±0.17)	49.56	38.83	42.39	0.20	72.31	0.39	5
RF	80.00% (±0.19)	72.24	66.21	67.70	0.09	85.81	0.73	2
NB	66.14% (±0.17)	55.59	45.69	48.47	0.22	71.35	0.43	4
LR	70.74% (±0.10)	49.26	37.11	41.14	0.16	75.89	0.48	3
CNN	94.48% (±0.11)	94.48	94.48	94.48	0	1	1	1

transcriptome, methylome, and integrated data and examined their association with specific subtypes. To find the co-expression module of feature methylation sites, we mapped the methylation site to gene name and extracted the gene expression data to construct co-expression modules. To construct the co-expression modules, we determined the soft threshold, β (β = 4, 6, and 5 for transcriptome, methylome, and integrated data, respectively) based on scale independence and mean connectivity (**Supplementary Figure S1**). We then merged modules with similarities above 0.6 for all three types of data. Finally, the dynamic tree cut showed a gene cluster dendrogram containing 3, 6, 5 co-expression models in the features of transcriptome, methylome, and integrated data, respectively (**Figures 4A,E,I**). To understand the genotype–phenotype relationship, we generated the module–trait relationship plot. We found distinct patterns of association between modules and subtypes (**Figures 4B,F,J**). Results show that the blue module (**Figure 4B**) was significantly and positively associated with the proneural subtype (r = 0.53, p = 4E-11). In contrast, it was negatively associated with the mesenchymal (r = -0.73, p = 2E-23), and weakly correlated with the classical subtype (r = 0.25, p = 0.004). Similarly, we found a distinct pattern of association between other modules (i.e., brown and turquoise) and subtypes (**Figure 4B**). We observed the same in the features from the methylome and integrated data. In methylome (**Figure 4F**), the brown module significantly and positively associated with only the proneural subtype (r = 0.33, p = 0.02). The green module is positively associated with the classical (r = 0.32, p = 0.03) and negatively associated with the proneural (r = -0.46, p = 9E-04). The blue module is strongly and positively correlated with the mesenchymal subtype (r = 0.55, p = 4E-05), whereas it was negatively associated with proneural (r = -0.6, p = 5E-06). However, the feature from the integrated data showed a more specific module–subtype association. At least one module was strongly and positively correlated with a specific subtype. The red (r = 0.64, p = 3E-07), turquoise (r = 0.66, p = 8E-08), and blue (r = 0.56, p = 1E-05) were explicitly and positively associated with classical, mesenchymal, and proneural, respectively (**Figure 4J**). The module–trait relationship analysis indicates that integration of transcriptome and methylome results in subsets of features strongly correlated with a particular subtype of GBM. Probably, the integrated data sets are mechanistically more relevant as the methylation and gene expression are integrated cellular processes. Next, we performed the gene set enrichment analysis (GSEA), i.e., GO biological process (BP) and molecular function (MF), using Enrichr to understand the biological relevance of each data type's top three positively correlated modules (Mallick et al., 2020). We observed that modules were significantly (adjusted p < .05) associated with several BP and MF that are linked to oncogenesis. For example, the turquoise module from the transcriptome data in the classical subtype is involved in the RIG-I signaling pathway that elicits RIG-I-like receptors' expression and activity (RLRs) (**Figure 4C**). These receptors stimulate both innate and adaptive immune responses against tumor antigens and promote the apoptosis of cancer cells (Bufalieri et al., 2021). In contrast, the brown module associated with the mesenchymal subtype (leukocyte adhesion to vascular endothelial cell) may be linked to the GBM-associated with the endothelial cell, that is, resistant to cytotoxic drugs, and also less apoptotic than healthy cells (Charalambous et al., 2006) (**Figure 4C**). Phosphatidylinositol 3 phosphate activity enriched in the turquoise module, solute proton symporter activity in the brown module, and syndecan binding in the blue module are associated with higher tumor grades and poor prognosis in GBM (Shi et al., 2017) (**Figure 4D**). Similarly, we observed that the blue module in the mesenchymal

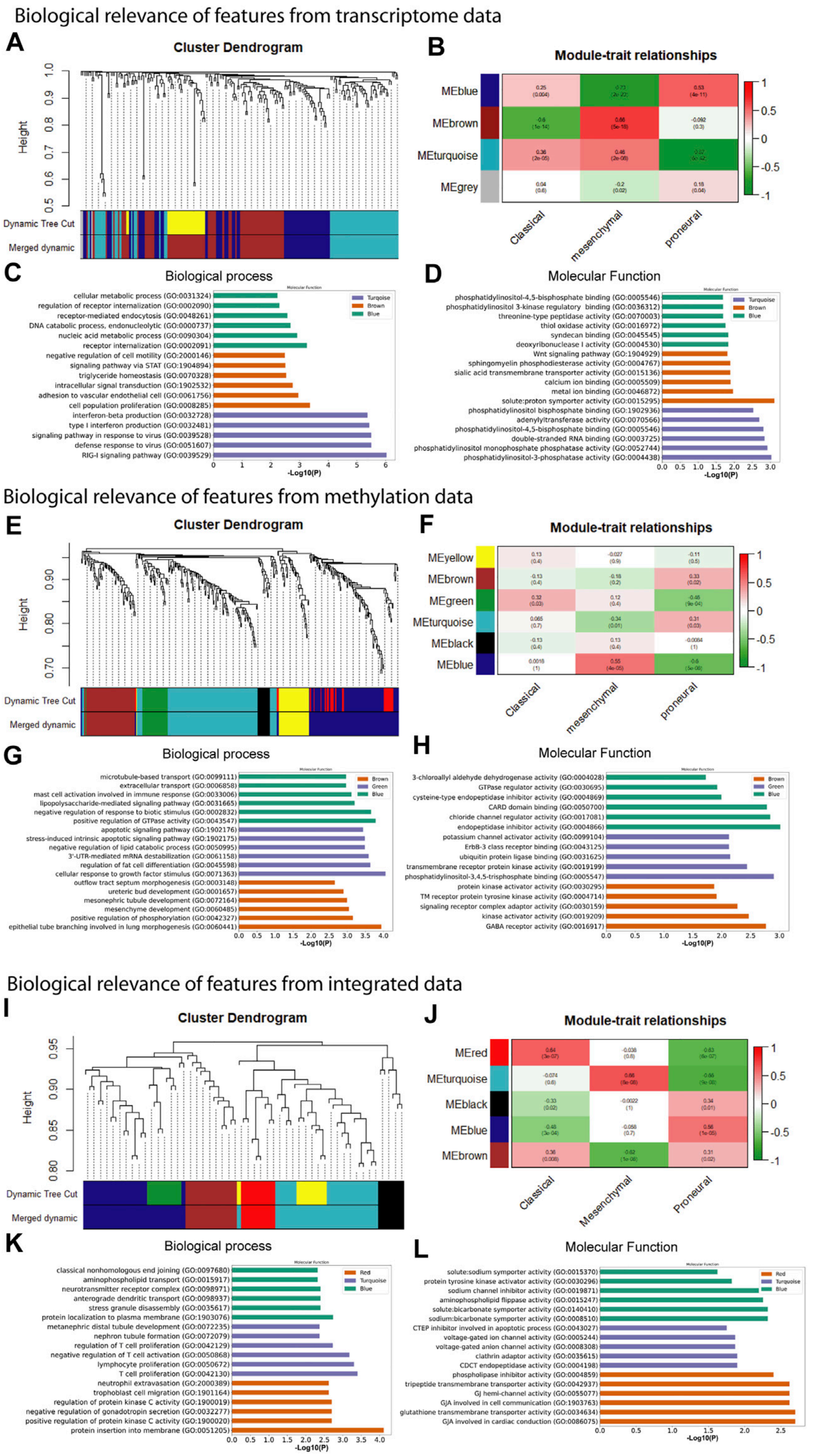

FIGURE 4 | Weighted gene co-expression network analysis and gene set enrichment of feature used for model building. **(A)** co-expression gene module, **(B)** module-trait relationship, **(C)** biological process, and **(D)** molecular function of feature from transcriptome data. **(E)** co-expression gene module, **(F)** module-trait relationship, **(G)** biological process, and **(H)** molecular function of feature from methylome data. **(I)** co-expression gene module, **(J)** module-trait relationship, **(K)** GO biological process term analysis, and **(L)** GO molecular function of feature from integrated data.

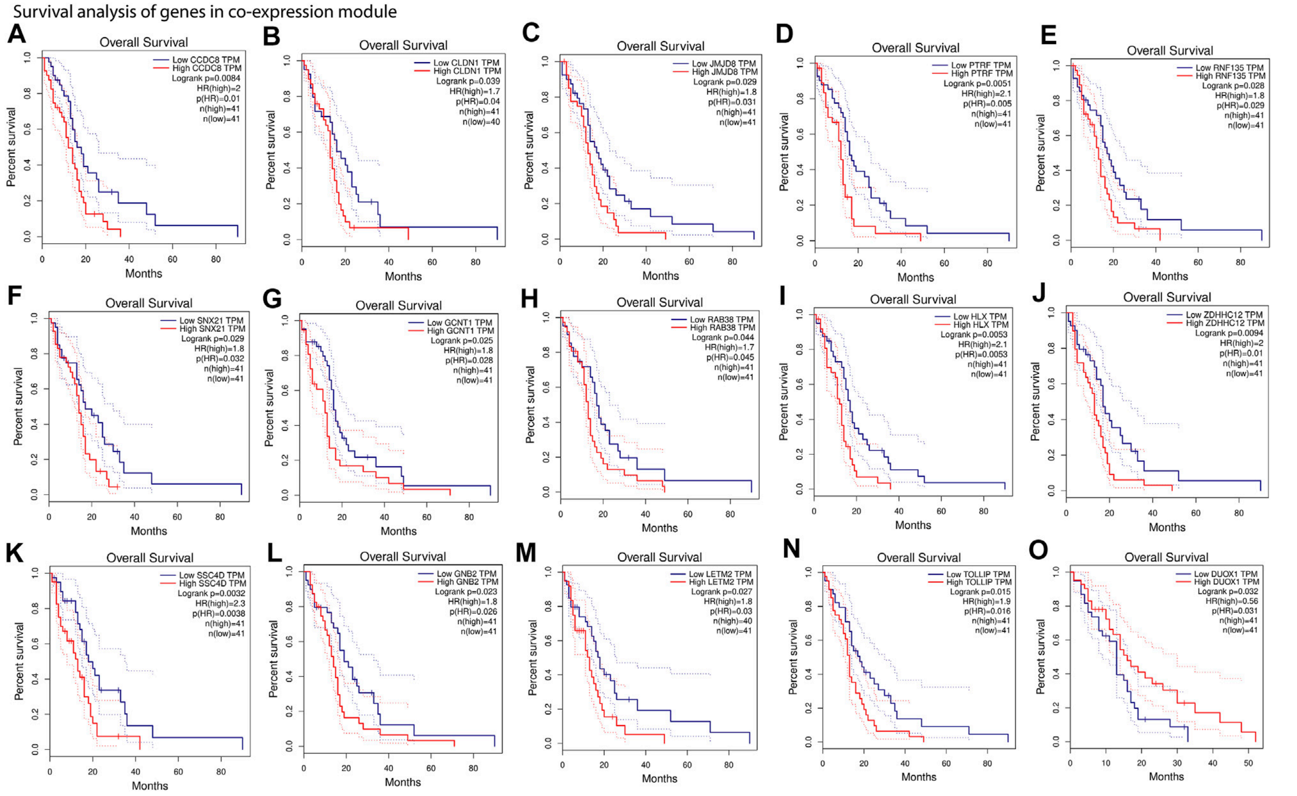

FIGURE 5 | Survival analysis of gene present in co-expression module. **(A–F)**, Kaplan–Meier plots of genes from positively associated modules with the classical subtype. **(G–M)**, Kaplan–Meier plots of genes from positively associated modules with the mesenchymal subtype. **(N,O)**, Kaplan–Meier plots of genes from positively associated modules with the proneural subtype. Overall survival was analyzed based on the clinical information of the patients from TCGA and quartile method of 75% cutoff of higher and 25% cutoff of lower limit (An extended version of this figure is provided in **Supplementary Figure S3**).

and the brown module in the proneural are linked to positive regulation of GTPase activity and positive regulation of phosphorylation in methylome data (**Figure 4G**). These processes are signatures of GBM formation and progression (He et al., 2021). Even molecular functions of several co-expression modules are involved in tumorigenesis, such as phosphatidylinositol 3, 4, 5 triphosphate binding enriched in the green module deregulates many key signaling pathways involving growth, proliferation, survival, and apoptosis in GBM (Mao et al., 2012) (**Figure 4H**). Furthermore, endopeptidase inhibitor activity, GABA receptor activity enriched in blue and brown modules, respectively, are predominant events in GBM (Labrakakis et al., 1998; Lin et al., 2020) (**Figure 4H**). The gene co-expressed modules in the integrated data, i.e., and the turquoise module (mesenchymal) involved with negative regulation of T cell activation and proliferation is one of the signatures of GBM (Woroniecka et al., 2018). The MF of the same module shows it is associated with gap junction channel activity involved in cell communication, which is also linked to GBM (Aasen et al., 2016) (**Figures 4K,L**).

Our results show that most of the positively correlated modules in GBM subtypes were involved in several BP and MF. Besides this, many of these BP and MF are involved in oncogenic processes. This shows a possibility of identifying these modules' genes as cancer biomarkers for therapy or diagnosis. We performed survival analysis of positively correlated modules (**Supplementary Figure S2**). The turquoise module in the integrated feature is significantly (log-rank test, p = .029) associated with the patient survival. Hence, we performed survival analysis of all genes separately present in these modules using GEPIA web tools (**Figure 5** and **Supplementary Figure S3**). We found several genes that were present in the co-expression module and also associated with the patient's survival (log-rank test, p < .05). The higher expression of most of the genes was associated with worse survival of the patients, except DUOX1 (FIGURE 5O) and FOXN2 (**Supplementary Figure S3**). However, higher or lower expression of genes associated with worse survival can be considered biomarkers (Sun et al., 2019; Liu et al., 2021). Furthermore, several experimental articles confirm the involvement of these genes in GBM formation and progression. For example, CCDC8, CLDN1, JMJD8, PTRF, RNF135, and SNX21 in classical (Berezovsky et al., 2014; Karnati et al., 2014; Pangeni et al., 2015; Yeo et al., 2016; Huang et al., 2018; Zhang et al., 2019) (**Figure 5A–F**); GCNT1, RAB38, HLX, ZDHHC12, SRCRB4D (SSC4D), GNB2, and LETM2 in mesenchymal (Thaker et al., 2009; Chen et al., 2014; Toton et al., 2018; Chen et al., 2020; Bianchetti et al., 2021; Giambra et al., 2021; Katsushima et al., 2021) (**Figure 5G–M**); and TOLLIP and DUOX1 (Humbert-Claude et al., 2016; Little et al., 2016) in proneural (**Figure 5N–O**) are linked to GBM patient survival. The association of genes from the modules with patient survival shows the possibility to identify them as subtype-specific prognostic biomarkers. We also observed that the expression pattern of survival-associated genes varied across the subtype (**Supplementary Figure S4**). Furthermore, we illustrated with gene enrichment analysis that their biological process and molecular functions are also linked to oncogenic events. Therefore, these findings confirm the clinical validity of our models and can provide insight into the complex regulatory processes in different subtypes of GBM.

DISCUSSION

The present study indicates that DL and ML can be powerful tools for finding patterns in large-scale genetic and epigenetic data sets related to human cancer. In general, efficient DL and ML tools work like a black box; researchers or clinicians may not be confident in diagnosing or classifying cancer patients using these approaches. However, if the basis of classification is biologically relevant and has higher accuracy, the diagnosis and patient management are more assured and systematic. Here, we present a biologically relevant DL- and ML-based framework to classify the subtype of GBM to increase accuracy in diagnosis; in turn, it can lead to better patient management. Previous studies try to develop the cancer classification model using a single type of omics data. Models are mainly developed for binary classification to identify healthy and cancer patients. However, we use two types of high-throughput data, i.e., transcriptome and methylome; integrated forms of these data were explored to develop the classification framework. Most importantly, we successfully separate three subtypes, classical, mesenchymal, and proneural, of GBM. Although we dealt with multiclass classification problems, we still achieved classification accuracy >90%. We also compared DL and ML techniques to identify the most suitable method for interpreting the transcriptome, methylome, and integrated data. The DL method, i.e., CNN, outperforms other ML models. Using CNN, we were able to classify the tumor into the correct subtype from the test and external cohort. We observed that overall classification performance was higher using the transcriptome and integrated data than the methylome data.

Another significant aspect of our findings is the biological relevance of features and the identification of subtype-specific prognostic biomarkers. To find the association of feature genes with specific subtypes, we performed WGCNA. The gene co-expression module-subtype relation analysis revealed how a subset of features is strongly and positively correlated with a particular subtype of GBM. In addition to that, the gene set enrichment analysis revealed that all positively correlated modules are biologically relevant even those that are linked to oncogenic processes. Among all data types, a strong module–trait relationship was observed in feature genes from integrated data. Furthermore, we identified several genes present in these co-expressed modules, which were linked to patient survival. Our study explained how the feature genes from the DL/ML framework could be used to find the subtype-specific biomarkers. Good agreement was found when comparing prognostic markers from this work against published experimental data. The feature genes of this study and CNN can provide assured and clinically relevant deep learning-based diagnostic tools for the proper treatment of GBM patients. Furthermore, the results of this work unravel and shed light on the understanding of genotype-phenotype relationships of the GBM subtype. Last, much of the research presented in this work can be applied to other human cancers to design DL-based diagnostic tools using high-throughput experimental data.

DATA AVAILABILITY STATEMENT

Publicly available data sets were analyzed in this study. This data can be found here: Cancer patient transcriptome data is available at https://xenabrowser.net/datapages/?dataset=TCGA.GBM.sampleMap%2FHiSeqV2&amp;host=https%3A%2F%2Ftcga.xenahubs.net&amp;removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443, and methylome data is available at https://xenabrowser.net/datapages/?dataset=TCGA.GBM.sampleMap%2FHumanMethylation450&amp;host=https%3A%2F%2Ftcga.xenahubs.net&amp;removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443. The validation data GSE145645 is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145645, and GSE128654 is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128654.

ETHICS STATEMENT

This article does not contain any studies with human participants or animals performed by any of the authors.

AUTHOR CONTRIBUTIONS

SM performed the experiment. TS and SM contributed and verified the code. ABD and ZZ conceived and planned the experiments. ABD wrote the manuscript with the input from all authors All authors discussed the results and contributed to the final manuscript.

FUNDING

Z.Z. was partially supported by the Cancer Prevention and Research Institute of Texas (CPRIT RP180734). A.B.D thanks National Institute of Technology, at Warangal India, for computational facility and fellowship for Ph.D. student.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.855420/full#supplementary-material

REFERENCES

Aasen, T., Mesnil, M., Naus, C. C., Lampe, P. D., and Laird, D. W. (2016). Gap Junctions and Cancer: Communicating for 50 Years. *Nat. Rev. Cancer* 16 (12), 775–788. doi:10.1038/nrc.2016.105

Afifi, S., GholamHosseini, H., and Sinha, R. (2017). "SVM Classifier on Chip for Melanoma Detection," in Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2017, Jeju, Korea (South), 11-15 July 2017, 270–274. doi:10.1109/EMBC.2017.8036814

Aran, D., Sabato, S., and Hellman, A. (2013). DNA Methylation of Distal Regulatory Sites Characterizes Dysregulation of Cancer Genes. *Genome Biol.* 14 (3), R21. doi:10.1186/GB-2013-14-3-R21

Bandyopadhyay, S., Mallik, S., and Mukhopadhyay, A. (2014). A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. *Ieee/acm Trans. Comput. Biol. Bioinf.* 11, 95–115. doi:10.1109/tcbb.2013.147

Basu, A., and Tiwari, V. K. (2021). Epigenetic Reprogramming of Cell Identity: Lessons from Development for Regenerative Medicine. *Clin. Epigenet* 13 (1), 144. doi:10.1186/S13148-021-01131-4

Berezovsky, A. D., Poisson, L. M., Cherba, D., Webb, C. P., Transou, A. D., Lemke, N. W., et al. (2014). Sox2 Promotes Malignancy in Glioblastoma by Regulating Plasticity and Astrocytic Differentiation. *Neoplasia* 16 (3), 193–206. doi:10.1016/j.neo.2014.03.006

Bianchetti, E., Bates, S. J., Nguyen, T. T. T., Siegelin, M. D., and Roth, K. A. (2021). RAB38 Facilitates Energy Metabolism and Counteracts Cell Death in Glioblastoma Cells. *Cells* 10 (7), 1643. doi:10.3390/CELLS10071643

Bozdag, S., Li, A., Riddick, G., Kotliarov, Y., Baysan, M., Iwamoto, F. M., et al. (2013). Age-specific Signatures of Glioblastoma at the Genomic, Genetic, and Epigenetic Levels. *PLoS ONE* 8, e62982. doi:10.1371/JOURNAL.PONE.0062982

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The Somatic Genomic Landscape of Glioblastoma. *Cell* 155 (2), 462–477. doi:10.1016/j.cell.2013.09.034

Bufalieri, F., Basili, I., di Marcotullio, L., and Infante, P. (2021). Harnessing the Activation of RIG-I Like Receptors to Inhibit Glioblastoma Tumorigenesis. *Front. Mol. Neurosci.* 14, 710171. doi:10.3389/fnmol.2021.710171

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., et al. (2020). Pan-Cancer Analysis of Whole Genomes. *Nature* 578 (7793), 82–93. doi:10.1038/s41586-020-1969-6

Chakravarthi, B. V. S. K., Nepal, S., and Varambally, S. (2016). Genomic and Epigenomic Alterations in Cancer. *Am. J. Pathol.* 186 (7), 1724–1735. doi:10.1016/j.ajpath.2016.02.023

Charalambous, C., Chen, T. C., and Hofman, F. M. (2006). Characteristics of Tumor-Associated Endothelial Cells Derived from Glioblastoma Multiforme. *Neurosur. Focus* 20 (4), E22. doi:10.3171/foc.2006.20.4.e22

Chen, X., Li, H., Fan, X., Zhao, C., Ye, K., Zhao, Z., et al. (2020). Protein Palmitoylation Regulates Cell Survival by Modulating XBP1 Activity in Glioblastoma Multiforme. *Mol. Ther. - Oncolytics* 17, 518–530. doi:10.1016/j.omto.2020.05.007

Chen, Z., Gulzar, Z. G., St. Hill, C. A., Walcheck, B., and Brooks, J. D. (2014). Increased Expression of GCNT1 Is Associated with Altered O -Glycosylation of PSA, PAP, and MUC1 in Human Prostate Cancers. *Prostate.* 74 (10), 1059–1067. doi:10.1002/PROS.22826

Dhar, G. A., Saha, S., Mitra, P., and Nag Chaudhuri, R. (2021). DNA Methylation and Regulation of Gene Expression: Guardian of Our Health. *Nucleus.* 64, 259–270. doi:10.1007/S13237-021-00367-Y

Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). Gene Selection and Classification of Microarray Data Using Random forest. *BMC Bioinformatics* 7, 3. doi:10.1186/1471-2105-7-3

Dong, Z., and Cui, H. (2019). Epigenetic Modulation of Metabolism in Glioblastoma. *Semin. Cancer Biol.* 57, 45–51. doi:10.1016/j.semcancer.2018.09.002

Giambra, M., Messuti, E., di Cristofori, A., Cavandoli, C., Bruno, R., Buonanno, R., et al. (2021). Characterizing the Genomic Profile in High-Grade Gliomas: From Tumor Core to Peritumoral Brain Zone, Passing through Glioma-Derived Tumorspheres. *Biology* 10 (11), 1157. doi:10.3390/BIOLOGY10111157

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and Interpreting Cancer Genomics Data via the Xena Platform. *Nat. Biotechnol.* 38 (6), 675–678. doi:10.1038/S41587-020-0546-8

He, H., Huang, J., Wu, S., Jiang, S., Liang, L., Liu, Y., et al. (2021). The Roles of GTPase-Activating Proteins in Regulated Cell Death and Tumor Immunity. *J. Hematol. Oncol.* 14 (1), 171. doi:10.1186/S13045-021-01184-1

Huang, K., Fang, C., Yi, K., Liu, X., Qi, H., Tan, Y., et al. (2018). The Role of PTRF/Cavin1 as a Biomarker in Both Glioma and Serum Exosomes. *Theranostics.* 8, 1540–1557. doi:10.7150/THNO.22952

Humbert-Claude, M., Duc, D., Dwir, D., Thieren, L., Sandström von Tobel, J., Begka, C., et al. (2016). Tollip, an Early Regulator of the Acute Inflammatory Response in the Substantia Nigra. *J. Neuroinflammation.* 13 (1), 303. doi:10.1186/S12974-016-0766-5

Jayanthi, V. S. P. K. S. A., Das, A. B., and Saxena, U. (2020). Grade-Specific Diagnostic and Prognostic Biomarkers in Breast Cancer. *Genomics.* 112 (1), 388–396. doi:10.1016/j.ygeno.2019.03.001

Jurmeister, P., Bockmayr, M., Seegerer, P., Bockmayr, T., Treue, D., Montavon, G., et al. (2019). Machine Learning Analysis of DNA Methylation Profiles Distinguishes Primary Lung Squamous Cell Carcinomas from Head and Neck Metastases. *Sci. Transl. Med.* 11 (509), eaaw8513. doi:10.1126/SCITRANSLMED.AAW8513

Karnati, H., Panigrahi, M., Shaik, N., Greig, N., Bagadi, S., Kamal, M., et al. (2014). Down Regulated Expression of Claudin-1 and Claudin-5 and up Regulation of β-Catenin: Association with Human Glioma Progression. *CNS Neurol. Disord. Drug Targets* 13 (8), 1413–1426. doi:10.2174/1871527313666141023121550

Katsushima, K., Lee, B., Kunhiraman, H., Zhong, C., Murad, R., Yin, J., et al. (2021). The Long Noncoding RNA Lnc-HLX-2-7 Is Oncogenic in Group 3 Medulloblastomas. *Neuro. Oncol.* 23 (4), 572–585. doi:10.1093/NEUONC/NOAA235

Kaviarasi, R., and Gandhi Raj, R. (2019). Accuracy Enhanced Lung Cancer Prognosis for Improving Patient Survivability Using Proposed Gaussian Classifier System. *J. Med. Syst.* 43 (7), 201. doi:10.1007/s10916-019-1297-2

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Res.* 44 (W1), W90–W97. doi:10.1093/NAR/GKW377

Labrakakis, C., Patt, S., Hartmann, J., and Kettenmann, H. (1998). Functional GABA(A) Receptors on Human Glioma Cells. *Eur. J. Neurosci.* 10 (1), 231–238. doi:10.1046/J.1460-9568.1998.00036.X

Langfelder, P., and Horvath, S. (2008). WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics.* 9, 559. doi:10.1186/1471-2105-9-559

Lin, Y., Liao, K., Miao, Y., Qian, Z., Fang, Z., Yang, X., et al. (2020). Role of Asparagine Endopeptidase in Mediating Wild-Type P53 Inactivation of Glioblastoma. *J. Natl. Cancer Inst.* 112 (4), 343–355. doi:10.1093/JNCI/DJZ155

Little, A. C., Sham, D., Hristova, M., Danyal, K., Heppner, D. E., Bauer, R. A., et al. (2016). DUOX1 Silencing in Lung Cancer Promotes EMT, Cancer Stem Cell Characteristics and Invasive Properties. *Oncogenesis.* 5 (10), e261. doi:10.1038/oncsis.2016.61

Liu, Z., Bensmail, H., and Tan, M. (2012). Efficient Feature Selection and Multiclass Classification with Integrated Instance and Model Based Learning. *Evol. Bioinform. Online* 8, 197–205. doi:10.4137/EBO.S9407

Liu, Z., Ru, L., and Ma, Z. (2021). Low Expression of ADCY4 Predicts Worse Survival of Lung Squamous Cell Carcinoma Based on Integrated Analysis and Immunohistochemical Verification. *Front. Oncol.* 11, 2241. doi:10.3389/fonc.2021.637733

Locke, W. J., Guanzon, D., Ma, C., Liew, Y. J., Duesing, K. R., Fung, K. Y. C., et al. (2019). DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Front. Genet.* 10, 1150. doi:10.3389/fgene.2019.01150

Maegawa, S., Hinkal, G., Kim, H. S., Shen, L., Zhang, L., Zhang, J., et al. (2010). Widespread and Tissue Specific Age-Related DNA Methylation Changes in Mice. *Genome Res.* 20 (3), 332–340. doi:10.1101/gr.096826.109

Mallick, K., Mallik, S., Bandyopadhyay, S., and Chakraborty, S. (2020). A Novel Graph Topology Based GO-Similarity Measure for Signature Detection from Multi-Omics Data and its Application to Other Problems. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 1. doi:10.1109/tcbb.2020.3020537

Mallik, S., Qin, G., Jia, P., and Zhao, Z. (2020a). Molecular Signatures Identified by Integrating Gene Expression and Methylation in Non-Seminoma and

Seminoma of Testicular Germ Cell Tumours. *Epigenetics.* 16 (2), 162–176. doi:10.1080/15592294.2020.1790108

Mallik, S., Seth, S., Bhadra, T., and Zhao, Z. (2020b). A Linear Regression and Deep Learning Approach for Detecting Reliable Genetic Alterations in Cancer Using DNA Methylation and Gene Expression Data. *Genes* 11 (8), 931. doi:10.3390/GENES11080931

Mallik, S., and Zhao, Z. (2020). Graph- and Rule-Based Learning Algorithms: a Comprehensive Review of Their Applications for Cancer Type Classification and Prognosis Using Genomic Data. *Brief Bioinform.* 21 (2), 368–394. doi:10.1093/BIB/BBY120

Mao, H., Lebrun, D. G., Yang, J., Zhu, V. F., and Li, M. (2012). Deregulated Signaling Pathways in Glioblastoma Multiforme: Molecular Mechanisms and Therapeutic Targets. *Cancer Invest.* 30 (1), 48–56. doi:10.3109/07357907.2011.630050

Mardis, E. R., and Wilson, R. K. (2009). Cancer Genome Sequencing: A Review. *Hum. Mol. Genet.* 18 (R2), R163–R168. doi:10.1093/HMG/DDP396

Moore, L. D., Le, T., and Fan, G. (2012). DNA Methylation and its Basic Function. *Neuropsychopharmacol* 38 (1), 23–38. doi:10.1038/npp.2012.112

Muthukrishnan, R., and Rohini, R. (2016). "LASSO: A Feature Selection Technique in Predictive Modeling for Machine Learning," in 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24-24 Oct. 2016, 18–20. doi:10.1109/ICACA.2016.7887916

Pangeni, R. P., Channathodiyil, P., Huen, D. S., Eagles, L. W., Johal, B. K., Pasha, D., et al. (2015). The GALNT9, BNC1 and CCDC8 Genes Are Frequently Epigenetically Dysregulated in Breast Tumours that Metastasise to the Brain. *Clin. Epigenet* 7 (1), 57. doi:10.1186/S13148-015-0089-X

Park, A. K., Kim, P., Ballester, L. Y., Esquenazi, Y., and Zhao, Z. (2019). Subtype-specific Signaling Pathways and Genomic Aberrations Associated with Prognosis of Glioblastoma. *Neuro Oncol.* 21 (1), 59–70. doi:10.1093/NEUONC/NOY120

Qin, G., Mallik, S., Mitra, R., Li, A., Jia, P., Eischen, C. M., et al. (2020). MicroRNA and Transcription Factor Co-regulatory Networks and Subtype Classification of Seminoma and Non-seminoma in Testicular Germ Cell Tumors. *Sci. Rep.* 10 (1), 852. doi:10.1038/s41598-020-57834-w

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., et al. (2011). Validation of a DNA Methylation Microarray for 450,000 CpG Sites in the Human Genome. *Epigenetics* 6 (6), 692–702. doi:10.4161/EPI.6.6.16196

Shi, S., Zhong, D., Xiao, Y., Wang, B., Wang, W., Zhang, F. a., et al. (2017). Syndecan-1 Knockdown Inhibits Glioma Cell Proliferation and Invasion by Deregulating a C-src/FAK-Associated Signaling Pathway. *Oncotarget* 8 (25), 40922–40934. doi:10.18632/ONCOTARGET.16733

Si, T., Miranda, P., Galdino, J. V., and Nascimento, A. (2021). Grammar-Based Automatic Programming for Medical Data Classification: An Experimental Study. *Artif. Intell. Rev.* 54, 4097–4135. doi:10.1007/S10462-020-09949-9

Sumithra, B., Saxena, U., and Das, A. B. (2019). A Comprehensive Study on Genome-wide Coexpression Network of KHDRBS1/Sam68 Reveals its Cancer and Patient-Specific Association. *Sci. Rep.* 9 (1), 11083. doi:10.1038/s41598-019-47558-x

Sun, J., Long, Y., Peng, X., Xiao, D., Zhou, J., Tao, Y., et al. (2019). The Survival Analysis and Oncogenic Effects of CFP1 and 14-3-3 Expression on Gastric Cancer. *Cancer Cel Int* 19, 1–12. doi:10.1186/S12935-019-0946-3

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: A Web Server for Cancer and normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res.* 45 (W1), W98–W102. doi:10.1093/NAR/GKX247

Thaker, N. G., Zhang, F., McDonald, P. R., Shun, T. Y., Lewen, M. D., Pollack, I. F., et al. (2009). Identification of Survival Genes in Human Glioblastoma Cells by Small Interfering RNA Screening. *Mol. Pharmacol.* 76 (6), 1246–1255. doi:10.1124/MOL.109.058024

Toton, E., Romaniuk, A., Konieczna, N., Hofmann, J., Barciszewski, J., and Rybczynska, M. (2018). Impact of PKCε Downregulation on Autophagy in Glioblastoma Cells. *BMC Cancer.* 18 (1), 185. doi:10.1186/S12885-018-4095-1

Triantaphyllou, E. (2000). *Multi-Criteria Decision Making Methods: A Comparative Study*. Boston, MA: Springer US. doi:10.1007/978-1-4757-3157-6

Van den Bent, M. J. (2010). Interobserver Variation of the Histopathological Diagnosis in Clinical Trials on Glioma: A Clinician's Perspective. *Acta Neuropathol.* 120 (3), 297–304. doi:10.1007/S00401-010-0725-7

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell.* 17 (1), 98–110. doi:10.1016/j.ccr.2009.12.020

Vinel, C., Rosser, G., Guglielmi, L., Constantinou, M., Pomella, N., Zhang, X., et al. (2021). Comparative Epigenetic Analysis of Tumour Initiating Cells and Syngeneic EPSC-Derived Neural Stem Cells in Glioblastoma. *Nat. Commun.* 12 (1), 6130. doi:10.1038/S41467-021-26297-6

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., et al. (2018). Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell.* 33 (1), 152. doi:10.1016/j.ccell.2017.12.012

Wang, X., Li, Y., Hu, H., Zhou, F., Chen, J., and Zhang, D. (2020). Comprehensive Analysis of Gene Expression and DNA Methylation Data Identifies Potential Biomarkers and Functional Epigenetic Modules for Lung Adenocarcinoma. *Genet. Mol. Biol.* 43 (3), e20190164. doi:10.1590/1678-4685-GMB-2019-0164

Witthayanuwat, S., Pesee, M., Supaadirek, C., Supakalin, N., Thamronganantasakul, K., and Krusun, S. (2018). Survival Analysis of Glioblastoma Multiforme. *Asian Pac. J. Cancer Prev.* 19 (9), 2613–2617. doi:10.22034/APJCP.2018.19.9.2613

Woroniecka, K. I., Rhodin, K. E., Chongsathidkiet, P., Keith, K. A., and Fecci, P. E. (2018). T-cell Dysfunction in Glioblastoma: Applying a New Framework. *Clin. Cancer Res.* 24 (16), 3792–3802. doi:10.1158/1078-0432.CCR-18-0047

Yang, X., Han, H., De Carvalho, D. D., Lay, F. D., Jones, P. A., and Liang, G. (2014). Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell.* 26 (4), 577–590. doi:10.1016/j.ccr.2014.07.028

Yeo, K. S., Tan, M. C., Wong, W. Y., Loh, S. W., Lam, Y. L., Tan, C. L., et al. (2016). JMJD8 Is a Positive Regulator of TNF-Induced NF-κB Signaling. *Sci. Rep.* 6, 34125. doi:10.1038/SREP34125

Yoon, J., Kim, M., Posadas, E. M., Freedland, S. J., Liu, Y., Davicioni, E., et al. (2021). A Comparative Study of PCS and PAM50 Prostate Cancer Classification Schemes. *Prostate Cancer Prostatic Dis.* 24 (3), 733–742. doi:10.1038/s41391-021-00325-4

Zhang, P., Xia, Q., Liu, L., Li, S., and Dong, L. (2020). Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy. *Front. Mol. Biosci.* 7, 562798. doi:10.3389/fmolb.2020.562798

Zhang, Y., Sui, R., Chen, Y., Liang, H., Shi, J., and Piao, H. (2019). Downregulation of miR-485-3p Promotes Glioblastoma Cell Proliferation and Migration via Targeting RNF135. *Exp. Ther. Med.* 18 (1), 475–482. doi:10.3892/etm.2019.7600

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.