Molecular Subtyping of Diffuse Gliomas using Magnetic Resonance Imaging: Comparison and Correlation between Radiomics and Deep Learning

Authors: Yiming Li[1]*, Dong Wei[2]*, Xing Liu[3], Xing Fan[3], Kai Wang[4], Shaowu Li[3], Zhong Zhang[1], Tianyi Qian[5], Jia Chang[6], Tao Jiang[1,3,7,8,9]†, Yefeng Zheng[2]†, Yinyan Wang[1]†

Affiliations:

[1]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China;

[2]YouTu Lab, Tencent, Shenzhen, China

[3]Beijing Neurosurgical Institute, Capital Medical University, Beijing, China;

[4]Department of Nuclear Medicine, Beijing Tiantan Hospital, Capital Medical University, Beijing, China;

[5] Healthcare Lab, Sinovation Ventures AI Institute

[6]Tencent HealthCare Co. Ltd., Shenzhen, China

[7]Center of Brain Tumor, Beijing Institute for Brain Disorders, Beijing, China;

[8]China National Clinical Research Center for Neurological Diseases;

[9]Chinese Glioma Genome Atlas Network (CGGA) and Asian Glioma Genome Atlas Network (AGGA)

* Co-first authors

† Co-corresponding authors

One Sentence Summary: Both the radiomic and deep convolutional neural network models could preoperatively predict the molecular subtypes of diffuse gliomas, and the later outperformed the former in most circumstances.

Abstract Radiomics-based and deep learning-based predictive models were established based on preoperative MRI and their performances were compared. The retrospective study randomized 1,016 diffuse glioma patients into training (model construction; n = 780) and validation (n = 236) sets. In the 2016 WHO classification scheme, diffuse gliomas are classified with four binary classification tasks (I–IV). In the independent validation set, the accuracies of the DCNN models (0.74–0.83) outperformed the radiomic models in all tasks, and the AUCs of the DCNN models outperformed the radiomic models in tasks I, II, and III. In task IV, the AUC of the DCNN model (0.66) was lower than that of the radiomic model (0.68). Both the radiomic and DCNN models could preoperatively predict the molecular subtypes of diffuse gliomas, and the later outperformed the former in most circumstances. Thus, the deep learning method is very promising in facilitating presurgical diagnosis of molecular subtypes.

Introduction

Approximately 100,000 people are diagnosed with diffuse gliomas annually (*1*). Diffuse glioma is the most aggressive and malignant form of primary brain tumor, and has a high mortality rate (*2, 3*). In the 2007 World Health Organization (WHO) classification system, diffuse gliomas were divided into grades II–IV based on histological criteria (*4*). However, this classification scheme had high interobserver variability (*5, 6*), and survival can vary considerably, even within the same grade (*7*).

Recent advances in molecular biology have greatly improved the understanding of glioma pathogenesis. Two clinically significant genetic alterations, the mutation (mut) of the gene for isocitrate dehydrogenase (*IDH*) mut and the 1p19q codeletion status, were integrated into the revised 2016 WHO classification system for diffuse gliomas (*8, 9*). Diffuse gliomas now comprise five molecular subtypes: (I) oligodendroglioma or anaplastic oligodendroglioma, *IDH*-mut and 1p19q co-deleted; (II) diffuse or anaplastic astrocytoma, *IDH*-mut; (III) diffuse or anaplastic astrocytoma, *IDH*-wildtype (wt); (IV) glioblastoma (GBM), *IDH*-mut; and (V) GBM, *IDH*-wt. Subtypes I, II, and III are lower-grade gliomas (LGGs). The new classification system has substantially improved the prediction of patient outcomes and the guidance of individualized treatment (*10-13*).

Imaging is widely used in clinical practice to aid decision making and is an important aspect of medical science (*14*). The development of computational techniques has included radiomics (*15*). In radiomics, medical images with biological information of tumors are transformed into high-dimensional data that can be analyzed. This approach has been applied to clinical decision-making systems to improve diagnosis (*16*), treatment (*17*), and prognosis (*18-20*). Specifically, radiomics allows the prediction of IDH and 1p19q status of gliomas with accuracies between 87.7% and

96.1% (*21*). Thus, the radiomics-based method could be a reliable alternative to predict the subtypes of diffuse gliomas.

The success of deep learning for various applications has generated a great deal of interest. Convolutional neural networks (CNNs) are a form of neural network specifically designed for image data processing. Deep CNNs (DCNNs) use a stack of convolutional layers (from a few to more than 100 layers, hence "deep") to learn different levels of abstraction from the input image data. DCNNs have achieved state-of-the-art results in a wide range of visual tasks, including outstanding performances in numerous medical image analysis applications compared to traditional image processing and computer vision techniques (*22-24*). This improved performance requires sufficient training data. However, the capability of DCNNs for magnetic resonance imaging (MRI)-based diffuse glioma subtyping in accordance with the WHO 2016 classification system (*9*) remains unexplored.

In this study, based on the data from a large number of patients, we developed radiomics and DCNN predictive systems for the molecular subtyping of diffuse gliomas. We hypothesized that the DCNN-based method could perform as well and even better than the performance of the radiomics-based method. Additionally, correlations between the radiomic and DCNN features were comprehensively were investigated.

Results

Clinical characteristics

A total of 780 patients were randomly assigned to the training set (445 males, 335 females; median age 47 years), and 236 patients were randomly assigned to the validation cohort (132 males, 104 females; median age 47 years). No significant difference was found in age, sex, or molecular subtypes between the two sets of patients.

Detailed clinical characteristics of the patients are summarized in Table 1. The study design is depicted in Fig. 1.

Performance comparison between the radiomic and DCNN models

Training set The receiver operating characteristics (ROC) curves of the radiomic models on the five-fold cross-validation training set are shown in Supplementary Fig. 1A–D. The areas under the curve (AUCs) of the mean ROC curves ranged from 0.85 to 0.93. The ROC curves of the DCNN models on the five-fold cross-validation training set are shown in Supplementary Fig. 1E–H. AUCs of the mean ROC curves ranged from 0.74 to 0.91. The model performances on the entire training set are shown in Fig. 2A–C. The AUCs of the DCNN models (ranging from 0.99–1.00) outperformed both the radiomic models and radiomic ensemble models in all tasks, and the accuracies of the DCNN models (ranging from 0.90–0.94) outperformed both the radiomic models and radiomic ensemble models in tasks I, II, and III. In task IV, the accuracy of the DCNN model was 0.90, lower than that of the radiomic model (0.92). The detailed data about the AUC, accuracy, sensitivity and specificity of the models are shown in Table 2. We were fully aware that machine learning models often tend to overfit the training data. Thus, we were more concerned with the performance of our radiomic and DCNN models on the independent validation set.

Validation set The established radiomics-based and DCNN-based models were subsequently applied to an independent validation set. Notably, the cutoff values applied in the validation set were the same as those identified with the training set. The ROC curves of the radiomic models are illustrated in Fig. 2D and 2E (AUCs ranged from 0.67–0.84), and those of the DCNN models are illustrated in Fig. 2F (AUCs ranged from 0.66–0.89). The accuracies of the DCNN models (ranging from 0.74–0.83) outperformed both the radiomic models and radiomic ensemble models in all tasks, and

the AUCs of the DCNN models outperformed both the radiomic models and radiomic ensemble models in tasks I, II, and III. In task IV, the AUC of the DCNN model was 0.66, which was slightly lower than those of the radiomic models (0.68 and 0.67). The detailed data about the AUC, accuracy, sensitivity and specificity of the models are shown in Table 3.

Radiomic and DCNN feature visualization and comparison

The t-distributed stochastic neighbor embedding (t-SNE) visualizations of both the radiomic and DCNN features for the four classification tasks are shown in Fig. 3. Dot distributions of such visualizations can intuitively reflect how well the visualized features can separate different classes (represented by distinct colors). From Fig. 3, two observations can be derived: (1) the radiomic features selected by LASSO can better discriminate diffuse glioma subtypes than the total collection of radiomic features, and (2) the DCNN features provide superior discriminative capability compared to the radiomic features, resulting in almost linearly-separable clusters. Exceptions to the first observation are for tasks II and IV, for which the LASSO-selected radiomic features did not seem to improve on the total radiomic features. In addition, although the clusters appeared more separable in the visualization of the DCNN features than in those of the radiomic features (Fig. 3D), still a considerable ratio of cases of IDH mut may be misclassified by the DCNN features given the very limited amount of IDH mut cases in glioblastoma multiforme (GBM).

Radiomic and DCNN feature correlation

Results of the correlation analysis using the Spearman's correlation (represented by ρ) are presented in Table 4. For each task, two settings of correlation analyses (A and B). Setting A was between the LASSO-selected radiomic features and all 2,560 DCNN features. Setting B was between the LASSO-selected radiomic features and the top 128

DCNN features. For quantification, we counted the numbers of ρ-values with absolute values > 0.5, and considered the corresponding pairs of radiomic and DCNN features to be correlated. We further calculated the ratios of correlated feature pairs with respect to the total numbers of feature pairs. Overall, the ratios were low, ranging from 0.01% to 1.96%. In both settings, the highest ratios were for task I (1.77% and 1.96% for A and B, respectively), followed by tasks III (0.69% and 0.73%, respectively), II (0.06% and 0.05%, respectively), and IV (0.03% and 0.01%, respectively). The ratios appeared to be substantially higher for tasks I and III than for tasks II and VI. In addition, we investigated how many radiomic and DCNN features actually composed these correlated pairs (since a radiomic feature could be correlated with more than one DCNN feature, and vice versa). For a specific task, the radiomic features were correlated with almost constant percentages of DCNN features (with respect to the total numbers of DCNN features used) across settings.

The correlation results in setting B were visualized using Circos for each task (Fig. 4). For task III, for example, nine radiomic features were correlated with 18 DCNN features, comprising a total of 46 connections (Fig. 4C). It was obviously apparent that more radiomic features were correlated with DCNN features for tasks I and III than for tasks II and IV, as represented by much denser connections. A comprehensive list of the radiomic features that were correlated with the DCNN features in setting B is provided in Supplementary Table 1.

Discussion

Radiomics-based and DCNN-based models were established in this study on the basis of preoperative multiparametric MRI. The molecular subtypes of diffuse gliomas were effectively predicted with these two methods. The radiomics approach used the open

source platform “PyRadiomics” for feature extraction, the LASSO algorithm for feature selection, and the SVM classifier for model construction. The AUCs of the radiomic models ranged from 0.92 to 0.97 in the entire training set, and from 0.67 to 0.84 in the independent validation set. As for the DCNN models, we developed a 2.5D network employing the ResNet18 model as the backbone for the binary classification tasks of diffuse glioma subtyping. The AUCs of the DCNN models ranged from 0.99 to 1.00 in the entire training set, and from 0.66 to 0.89 in the independent validation set. Notably, the DCNN models outperformed the radiomic models in most circumstances. Additionally, the DCNN features demonstrated superior discriminative capability than the radiomic features in the t-SNE visualization analysis. Overall, our results indicate that both radiomics and deep learning could predict the molecular subtypes of diffuse gliomas non-invasively, and that deep learning performs better than radiomics when the study cohorts are large.

Our study expands the work of several recent studies that revealed novel associations between molecular status and MRI. Firstly, it was revealed that the IDH mutation status could be effectively predicted with multimodal MRI radiomic features (*25*) and residual convolutional neural network (*26*). A subsequent study revealed that the molecular subtypes of diffuse gliomas could be comprehensively predicted with radiomics (*21*). We used 1,016 patients to develop the predictive models, which is many more patients than in the previous studies. Such a (relatively) big patient cohort enabled, for the first time, a comprehensive prediction of the molecular subtypes of diffuse gliomas with DCNN models, which usually require a large number of cases for training to achieve reasonable performance. Additionally, performances and features of the different methodologies (radiomics and deep learning) were compared and correlated for the first time in predicting the 2016 WHO classification.

The observed performance of the machine learning models (including both the radiomic and DCNN models) agreed with the empirical experience of clinicians. That is, it is easier to identify GBM or LGGs than to identify the status of molecular markers based on multiparametric MRIs. Neither the radiomic or DCNN model were not satisfactory in differentiating between IDH mut and wt GBMs. The unsatisfactory performances might be caused by the relatively small sample size of IDH mut GBMs, and the corresponding serious imbalance between IDH mut and wt cases in GBMs. Nonetheless, looking at the two distinct clusters in the t-SNE visualization of the DCNN features for the specific task (Fig. 3D), we expect the performance of the DCNN model to improve with more and better-balanced training data. In addition, employing more sophisticated data augmentation techniques, such as generative adversarial neural networks (*27*), may improve the performance of the DCNN models.

The feature visualizations in Fig. 3 acquired using the t-SNE technique matched reasonably well with the prediction accuracies presented in Table 3. For the radiomic approach, the LASSO-selected features were more separable in the visualizations for the tasks of GBMs vs. LGGs, and 1p/19q codeletion vs. noncodeletion, than for the tasks of IDH mut vs. wt. Accordingly, the prediction accuracies were higher for the former two tasks than for the latter two. Taking one step further, the visualized DCNN features were much more separable than the radiomic features for all four tasks, corresponding to substantially higher AUCs and accuracies for the DCNN models. The feature correlation analysis (presented in Table 4 and Fig. 4) complemented the visual comparison analysis. Using 0.5 as the cutoff value, the radiomic features correlated with low ratios of the DCNN features for tasks II–IV, ranging from 0.78% to 14.06% (the exception was task I, with ratios above 50%), suggesting little overlap in the subtype-discriminative information expressed by these two groups of features. Results

from the feature comparison and correlation analyses could suggest that the collection of hand-crafted radiomic features computed from the data used in this study were insufficient for the subtyping of diffuse glioma (especially for the task of IDH mut vs. wt in LGGs) compared to the automatically learnt DCNN features.

Radiomic features are crafted generically by predefined formulations. Accordingly, researchers must generate a large pool of candidate radiomic features and then select a subset that gives the best performance for the specific tasks involved. When the candidate pool includes features that are discriminative for these tasks, the performance of radiomics-based approaches is competitive. However, being hand-crafted in advance, the radiomic features lack the flexibility in adaptation and thus may not always provide features that are sufficiently discriminative for particular tasks. In such cases, radiomics-based approaches would produce less satisfactory results. On the contrary, DCNNs are data-driven and able to automatically learn to extract task-specific features. Hence, when given enough data for effective training (e.g., for tasks I–III in this study), DCNN-based approaches are preferred to radiomics-based ones, because of their flexibility in adapting to different tasks.

The prediction of molecular subtypes is very important in the management of diffuse gliomas. Obtaining molecular subtype information preoperatively could guide the surgical strategy. A previous study revealed that surgical strategies for different IDH status should be different in patients with malignant diffuse astrocytoma. Both residual enhanced and non-enhanced tumor are associated with poor prognosis of IDH mut patients, while only residual enhanced tumor are associated with poor prognosis of IDH-wt patients (*28*), indicating that a larger resection extent should be suggested for IDH mut patients. Additionally, molecular targeted therapies have attracted a lot of attention (*8*) and neoadjuvant therapies could be developed in the future with the help

of preoperative diagnosis of molecular subtypes. Moreover, the 2016 WHO classification has significant prognostic value (*2, 13, 29*), and accurate preoperative prediction of molecular subtypes could help patients with poor prognosis participate in clinical trials in time.

The current study has a few limitations. Firstly, the study design was on the basis of retrospectively collected, single-institute imaging data. A prospective study that includes multi-center imaging data should be conducted to confirm our findings. Secondly, both the radiomics-based and DCNN-based approaches performed unsatisfactorily in task IV (IDH mut vs. wt in GBM). The small number of IDH mut GBMs and the resulting ill-balanced class distribution may be the main reasons. With increased data, task IV would be solved better in the future.

Conclusion

In this study, the machine learning models based on radiomics and deep learning were established. They could effectively predict the molecular subtypes of diffuse gliomas from preoperative multiparametric MRI data. The DCNN models outperformed the radiomic models in most circumstances, indicating the promising application of deep learning in molecular subtyping of diffuse gliomas. With the increased availability of additional data in the future, the DCNN-based approach has the potential to further performance improvement.

Materials and methods

Hierarchical diffuse glioma subtyping model

According to the 2016 WHO classification scheme, diffuse gliomas can be classified by a three-level subtyping model with four binary classification tasks (Fig.

5). At the top level, diffuse gliomas are classified into LGGs and GBM based on the histological phenotypes. At the middle level, both LGGs and GBMs can be classified as IDH mut or IDH-wt. At the bottom level, IDH mut in LGGs are further classified according to 1p19q codeletion or noncodeletion. This generic model for diffuse glioma subtyping has already been used in a previous study (*21*). We also adopted this model. Separate classifiers were built for the four binary classification tasks: I, Grading; II, LGGs_IDH; III, LGGs_IDH_1p19q; and IV, GBM_IDH. Compared to a single-level multiclass model, the three-level hierarchical model is more convenient to plug-in histologic and genotype information when available.

Study cohort

A total of 1,016 patients treated at the Beijing Tiantan Hospital from September 2014 to April 2018 were retrospectively analyzed. Inclusion criteria were: (1) pathologically diagnosed as primary diffuse gliomas; (2) available preoperative T1-weighted (T1w), T2-weighted (T2w), and T1 contrast enhancement (T1CE) MR images; (3) age ≥ 18 years; (4) available IDH status (detected with immunohistochemistry or pyrosequencing); and (5) available 1p19q status (detected using fluorescence in situ hybridization) for LGGs. A flowchart of patient inclusion/exclusion is provided in Supplementary Fig. 2. Ethical approval of this retrospective study was received from the institutional review board of the Beijing Tiantan Hospital.

The 1,016 patients were randomly (a case was either placed in the training set with 80% probability, or placed in the validation set with 20% probability) divided into the training set (n = 780) or validation set (n = 236). The training set was used for model construction and the validation set was used for model testing. Table 1 summarizes the clinical characteristics of the patients in the training and validation sets.

Image preprocessing and segmentation

Minimal preprocessing was performed. MRI volumes were resampled to the highest resolution (0.34×0.34×5 mm) among the images to avoid information loss during resampling. Tumor regions were delineated with T2w images, since identifying tumor borders of LGGs is hard in T1w and T1CE images. Therefore, the T1w and T1CE volumes were rigidly co-registered to the T2w volume. The tumor regions delineated in the T2w images were subsequently applied to corresponding T1w and T1CE images. The delineated tumor regions were used by both radiomics-based and DCNN-based methods for more focused feature extraction and less computational burden. Lastly, each MRI volume was normalized using z-score per instance; concretely, the z-score was calculated as $z = (x - \bar{x})/s$, where $\bar{x}$ and s are the sample mean and standard deviation, respectively.

The tumor regions were delineated on T2w images by an experienced neuroradiologist (J.M. with more than 15 years of experience) using ITK-snap software (http://www.itksnap.org) (*30*). Abnormal hyperintense signals (including edema) in the T2w images were considered as tumor regions, and the cerebrospinal fluid was avoided.

Radiomic model development

Radiomic feature extraction The open source "PyRadiomics" package (*31*) was used for extraction of MR radiomic features. In total, 3,362 radiomic features were extracted from the three MR sequences (T1w, T1CE, and T2w). A detailed description of the feature extraction process is provided in the Supplementary Material. The extracted radiomic features were normalized using z-score prior to subsequent steps.

Feature and model selection Overfitting is a practical concern with this many features. The least absolute shrinkage and selection operator (LASSO) (*32*) was employed to select a subset of radiomic features (patient age and sex were jointly selected with the radiomic features) that were predictive for each of the four binary classification tasks.

LASSO regression uses L1 regularization to simultaneously reduce overfitting and select features. A parameter designated α controls the extent of regularization: the larger α is, the fewer features are selected. We used five-fold cross-validation to pick the optimal α value from a pool of candidate values. This optimal value should: (1) give the best regression result and (2) ensure that the selected number of features was within 1/10 to 1/3 of the number of training samples, to avoid overfitting and underfitting (*33*). Next, a support vector machine (SVM) (*34, 35*) classifier was trained with the selected features. Again, five-fold cross-validation was employed to find the parameters of the optimal SVM model that yielded the highest mean AUC across folds in ROC curve analysis. The SVM parameters we considered include: (1) the kernel types, including the linear, radial-basis, and polynomial kernel functions, (2) the penalty parameter *C*, (3) the kernel coefficient γ, and (4) the degree of the polynomial function. After the optimal parameters were determined, a final SVM model was trained with these parameters on the entire training set.

Usually in a classical radiomics approach, a single final classifier is trained on the entire training data set after the feature and model selection, and applied to the independent validation set for evaluation. In this study, however, for a fair comparison with the DCNN approach, which adopted a five-model ensemble on the validation set, we also implemented an ensemble version of the radiomic approach. Concretely, five SVM-based classifiers were trained with the optimal parameters, each with four of the five folds of the training set. After that, test performance was evaluated by the ensemble of the five classifiers. This setting matched that of the DCNN approach described next.

DCNN model development

We developed a 2.5-dimensional (2.5D) DCNN model for the binary classification tasks of diffuse glioma subtyping, on the basis of the residual blocks (ResNet18 model)

for deep learning (*36*). The model structure is shown in Fig. 6 and a detailed description of the model is provided in the Supplementary Material. The models were trained on an NVIDIA Tesla P40 graphical processing unit using mini-batches of 32 cases. The binary cross entropy loss was used. The parameters of the models were optimized using the Adam algorithm (*37*) with a weight decay of 0.0005 for L2 regularization. The learning rate was initially set to 0.0005, and exponentially decayed by multiplying by the factor 0.97^{Ep}, with Ep being the epoch number.

For a given binary classification task, the training set was randomly divided into five folds, each of which was used as the validation fold in turn. When a specific fold was used for validation, the other four folds were used to train the DCNN model. The AUC was used as the stopping and model selection criteria. Concretely, the training was stopped 10 epochs after the training AUC rose above 0.99 for the first time, and the model with the highest validation AUC was kept. In addition, the optimal operating point was determined on the validation ROC curve where the true positive rate minus the false positive rate was maximal. Iterating through the five folds resulted in five individual DCNN models for the current classification task. At inference time, a test case was fed into all the five models to get five probabilities and five predictions. The final prediction probability was the mean of these five probabilities, and the final prediction was obtained by a majority voting.

Radiomic and DCNN feature visualization and comparison

We visualized and compared the radiomic and DCNN features for each of the four binary classification tasks. The 3,362 radiomic features extracted from the three MR sequences were normalized using the z-score. As to the DCNN features, the final, 512-dimensional feature vectors that were fed to the fully connected layers (i.e., fc layers) were pulled out from all the five DCNN models. Hence, the total number of DCNN

features is 512 × 5 = 2560. The radiomic and DCNN features of combined training and validation sets were subjected to the feature visualization analysis, since using the validation set alone would produce visualizations that were too sparse to clearly reflect underlying trends (Supplementary Fig. 3).

For a straightforward visual comparison of the radiomic and DCNN features, the t-SNE technique was employed. The t-SNE is a nonlinear technique for dimensionality reduction that is particularly well suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions (*38*). It has been extensively applied in a wide range of applications, especially for visualization of high-level representations learnt by machine learning models. In this study, we used t-SNE to visualize both the radiomic and DCNN features, for an intuitive perception of how well these features can distinguish diffuse gliomas of different subtypes.

Radiomic and DCNN feature correlation

In addition to the visual comparison, we correlated the radiomic and DCNN features for each of the four binary classification tasks using the Spearman's rank correlation coefficient. The features were extracted and prepared the same way as for the t-SNE visualization. Here, we only correlated the LASSO-selected radiomic features with the DCNN features, as only the LASSO-selected ones were actually used for classification by the radiomics approach. For visualization of the correlation analysis, the Circos software (http://circos.ca) was used (*39*). This software visualizes data in a circular layout, which makes it ideal for exploring relationships between objects. To avoid over-cluttering in the Circos visualization, we selected the top 128 DCNN features with largest absolute weights (in the fc layers of the DCNN models) for additional correlation analyses and visualization.

Statistical analyses

The statistical analyses were primarily conducted using the Python platform. Specifically, the radiomic feature extraction was conducted using the 'PyRadiomics' package; the LASSO regression, SVM classifier, ROC curve analysis, t-SNE feature visualization, and Spearman's correlation were conducted using the 'sklearn' package; and the CNN model was implemented using the 'PyTorch' package. Differences in clinical characteristics between the training and validation sets were evaluated using the Mann-Whitney U and chi-square tests, and a p-value < 0.05 was considered statistically significant. To evaluate and compare performances of the developed models, the metrics of the AUC and accuracy were employed.

List of Supplementary Materials

Supplementary material

Supplementary Fig. 1. Performances of the radiomic **(A–D)** and DCNN **(E–H)** models for the four binary classification tasks on the five-fold cross-validation training set. ROC curves for each fold and the mean ROC curves.

Supplementary Fig. 2. Work flow of patient inclusion and exclusion.

Supplementary Fig. 3. The t-SNE visualizations using the validation data alone for the four classification tasks **(A–D)** of the standardized total collection of radiomic features (including age and gender) prior to feature selection; standardized radiomic features selected by LASSO; and DCNN features as described in Method.

Supplementary Table 1. List of the radiomic features that were correlated with the DCNN features in Figure 4.

References

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424 (2018).
2. A. M. Molinaro, J. W. Taylor, J. K. Wiencke, M. R. Wrensch, Genetic and molecular epidemiology of adult diffuse glioma. *Nat Rev Neurol* **15**, 405-417 (2019).
3. T. Jiang, Y. Mao, W. Ma, Q. Mao, Y. You, X. Yang, C. Jiang, C. Kang, X. Li, L. Chen, X. Qiu, W. Wang, W. Li, Y. Yao, S. Li, S. Li, A. Wu, K. Sai, H. Bai, G. Li, B. Chen, K. Yao, X. Wei, X. Liu, Z. Zhang, Y. Dai, S. Lv, L. Wang, Z. Lin, J. Dong, G. Xu, X. Ma, J. Cai, W. Zhang, H. Wang, L. Chen, C. Zhang, P. Yang, W. Yan, Z. Liu, H. Hu, J. Chen, Y. Liu, Y. Yang, Z. Wang, Z. Wang, Y. Wang, G. You, L. Han, Z. Bao, Y. Liu, Y. Wang, X. Fan, S. Liu, X. Liu, Y. Wang, Q. Wang, G. Chinese Glioma Cooperative, CGCG clinical practice guidelines for the management of adult diffuse gliomas. *Cancer letters* **375**, 263-273 (2016).
4. D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, P. Kleihues, The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol* **114**, 97-109 (2007).
5. M. J. van den Bent, Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol* **120**, 297-304 (2010).
6. D. Sturm, B. A. Orr, U. H. Toprak, V. Hovestadt, D. T. W. Jones, D. Capper, M. Sill, I. Buchhalter, P. A. Northcott, I. Leis, M. Ryzhova, C. Koelsche, E. Pfaff, S. J. Allen, G. Balasubramanian, B. C. Worst, K. W. Pajtler, S. Brabetz, P. D. Johann, F. Sahm, J. Reimand, A. Mackay, D. M. Carvalho, M. Remke, J. J. Phillips, A. Perry, C. Cowdrey, R. Drissi, M. Fouladi, F. Giangaspero, M. Lastowska, W. Grajkowska, W. Scheurlen, T. Pietsch, C. Hagel, J. Gojo, D. Lotsch, W. Berger, I. Slavc, C. Haberler, A. Jouvet, S. Holm, S. Hofer, M. Prinz, C. Keohane, I. Fried, C. Mawrin, D. Scheie, B. C. Mobley, M. J. Schniederjan, M. Santi, A. M. Buccoliero, S. Dahiya, C. M. Kramm, A. O. von Bueren, K. von Hoff, S. Rutkowski, C. Herold-Mende, M. C. Fruhwald, T. Milde, M. Hasselblatt, P. Wesseling, J. Rossler, U. Schuller, M. Ebinger, J. Schittenhelm, S. Frank, R. Grobholz, I. Vajtai, V. Hans, R. Schneppenheim, K. Zitterbart, V. P. Collins, E. Aronica, P. Varlet, S. Puget, C. Dufour, J. Grill, D. Figarella-Branger, M. Wolter, M. U. Schuhmann, T. Shalaby, M. Grotzer, T. van Meter, C. M. Monoranu, J. Felsberg, G. Reifenberger, M. Snuderl, L. A. Forrester, J. Koster, R. Versteeg, R. Volckmann, P. van Sluis, S. Wolf, T. Mikkelsen, A. Gajjar, K. Aldape, A. S. Moore, M. D. Taylor, C. Jones, N. Jabado, M. A. Karajannis, R. Eils, M. Schlesner, P. Lichter, A. von Deimling, S. M. Pfister, D. W. Ellison, A. Korshunov, M. Kool, New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs. *Cell* **164**, 1060-1072 (2016).
7. A. L. Lin, L. M. DeAngelis, Reappraising the 2016 WHO classification for diffuse glioma. *Neuro Oncol* **19**, 609-610 (2017).
8. S. Lapointe, A. Perry, N. A. Butowski, Primary brain tumours in adults. *The Lancet* **392**, 432-446 (2018).
9. D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, D. W. Ellison, The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803-820 (2016).

10. G. Reifenberger, H. G. Wirsching, C. B. Knobbe-Thomsen, M. Weller, Advances in the molecular genetics of gliomas - implications for classification and therapy. *Nat Rev Clin Oncol* **14**, 434-452 (2017).
11. T. Iuchi, T. Sugiyama, M. Ohira, H. Kageyama, S. Yokoi, T. Sakaida, Y. Hasegawa, T. Setoguchi, M. Itami, Clinical significance of the 2016 WHO classification in Japanese patients with gliomas. *Brain Tumor Pathol* **35**, 71-80 (2018).
12. P. J. Cimino, M. Zager, L. McFerrin, H. G. Wirsching, H. Bolouri, B. Hentschel, A. von Deimling, D. Jones, G. Reifenberger, M. Weller, E. C. Holland, Multidimensional scaling of diffuse gliomas: application to the 2016 World Health Organization classification system with prognostically relevant molecular subtype discovery. *Acta Neuropathol Commun* **5**, 39 (2017).
13. E. Tabouret, A. T. Nguyen, C. Dehais, C. Carpentier, F. Ducray, A. Idbaih, K. Mokhtari, A. Jouvet, E. Uro-Coste, C. Colin, O. Chinot, H. Loiseau, E. Moyal, C. A. Maurage, M. Polivka, E. Lechapt-Zalcman, C. Desenclos, D. Meyronet, J. Y. Delattre, D. Figarella-Branger, P. N. For, Prognostic impact of the 2016 WHO classification of diffuse gliomas in the French POLA cohort. *Acta Neuropathol* **132**, 625-634 (2016).
14. H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **5**, 4006 (2014).
15. W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, R. H. Mak, R. M. Tamimi, C. M. Tempany, C. Swanton, U. Hoffmann, L. H. Schwartz, R. J. Gillies, R. Y. Huang, H. Aerts, Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin* **69**, 127-157 (2019).
16. Y.-q. Huang, C.-h. Liang, L. He, J. Tian, C.-s. Liang, X. Chen, Z.-l. Ma, Z.-y. Liu, Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. *Journal of Clinical Oncology* **34**, 2157-2164 (2016).
17. P. Kickingereder, M. Go tz, J. Muschelli, A. Wick, U. Neuberger, R. T. Shinohara, M. Sill, M. Nowosielski, H. P. Schlemmer, A. Radbruch, W. Wick, M. Bendszus, K. H. Maier-Hein, D. Bonekamp, Large-scale Radiomic Profiling of Recurrent Glioblastoma Identifies an Imaging Predictor for Stratifying Anti-Angiogenic Treatment Response. *Clinical Cancer Research* **22**, 5765-5771 (2016).
18. P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. Larue, A. J. G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* **14**, 749-762 (2017).
19. S. Napel, W. Mu, B. V. Jardim-Perassi, H. Aerts, R. J. Gillies, Quantitative imaging of cancer in the postgenomic era: Radio(geno)mics, deep learning, and habitats. *Cancer* **124**, 4633-4649 (2018).
20. S. Zhang, B. Zhang, J. Tian, D. Dong, D. S. Gu, Y. H. Dong, L. Zhang, Z. Y. Lian, J. Liu, X. N. Luo, S. F. Pei, X. K. Mo, W. H. Huang, F. S. Ouyang, B. L. Guo, L. Liang, W. Chen, C. H. Liang,

Radiomics features of Multiparametric MRI as Novel Prognostic Factors in Advanced Nasopharyngeal Carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*, (2017).

21. C. F. Lu, F. T. Hsu, K. L. Hsieh, Y. J. Kao, S. J. Cheng, J. B. Hsu, P. H. Tsai, R. J. Chen, C. C. Huang, Y. Yen, C. Y. Chen, Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas. *Clin Cancer Res* **24**, 4429-4436 (2018).
22. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, C. I. Sanchez, A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60-88 (2017).
23. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).
24. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster, Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402-2410 (2016).
25. B. Zhang, K. Chang, S. Ramkissoon, S. Tanguturi, W. L. Bi, D. A. Reardon, K. L. Ligon, B. M. Alexander, P. Y. Wen, R. Y. Huang, Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro Oncol* **19**, 109-117 (2017).
26. K. Chang, H. X. Bai, H. Zhou, C. Su, W. L. Bi, E. Agbodza, V. K. Kavouridis, J. T. Senders, A. Boaro, A. L. Beers, B. Zhang, A. Capellini, W. Liao, Q. Shen, X. Li, B. Xiao, J. Cryan, S. Ramkissoon, L. Ramkissoon, K. L. Ligon, P. Y. Wen, R. S. Bindra, J. H. Woo, O. Arnaout, E. Gerstner, P. J. Zhang, B. Rosen, L. Yang, R. Y. Huang, J. Kalpathy-Cramer, Residual Convolutional Neural Network for Determination of IDH Status in Low- and High-grade Gliomas from MR Imaging. *Clinical cancer research : an official journal of the American Association for Cancer Research*, (2017).
27. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, in *Advances in Neural Information Processing Systems*. (2014), pp. 2672-2680.
28. J. Beiko, D. Suki, K. R. Hess, B. D. Fox, V. Cheung, M. Cabral, N. Shonka, M. R. Gilbert, R. Sawaya, S. S. Prabhu, J. Weinberg, F. F. Lang, K. D. Aldape, E. P. Sulman, G. Rao, I. E. McCutcheon, D. P. Cahill, IDH1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection. *Neuro Oncol* **16**, 81-91 (2014).
29. H. G. Wirsching, M. Weller, The Role of Molecular Diagnostics in the Management of Patients with Gliomas. *Curr Treat Options Oncol* **17**, 51 (2016).
30. P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, G. Gerig, User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* **31**, 1116-1128 (2006).
31. J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, H. Aerts, Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* **77**, e104-e107 (2017).
32. R. Tibshirani, Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).
33. J. Hua, Z. Xiong, J. Lowey, E. Suh, E. R. Dougherty, Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* **21**, 1509-1515 (2005).

34. C.-W. Hsu, C.-C. Chang, C.-J. Lin, A practical guide to support vector classification. (2003).
35. G. Orru, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, A. Mechelli, Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews* **36**, 1140-1152 (2012).
36. K. He, Zhang, X., Ren, S. and Sun, J., Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778 (2016).
37. D. P. a. B. Kingma, J., Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, (2014).
38. G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, Y. Kluger, Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* **16**, 243-245 (2019).
39. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).

Funding: This study was supported by the National Natural Science Foundation of China (No. 81601452), the Beijing Natural Science Foundation (No. 7174295).

Author contributions: Data collection: Y.M.L., X.L., X.F., Z.Z.; data analyses: D.W., Y.M.L., Y.Y.W.; tumor segmentation: K.W., S.W.L.; study design: Y.Y.W., X.L., T.J.; manuscript writing: D.W., Y.M.L., Y.Y.W.; supervision: Y.Y.W., J.C., T.Y.Q. and Y.F.Z.

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: The raw data and codes could be provided when necessary through contacting the corresponding author.

Figure legends

Fig. 1. The radiomics and deep learning pipelines. These two classification systems were developed separately, and their performances were compared on an independent validation set. Abbreviations: LBP, local binary patterns; LASSO, least absolute shrinkage and selection operator; SVM, support vector machine; MP, max pooling; fc, fully connected; AUC, area under the curve.

Fig. 2. Prediction of subtypes of diffuse gliomas using the radiomic and DCNN models. **(A–C)** Receiver operating characteristic (ROC) curves of the four binary classification tasks for the radiomic and DCNN models on the training set. **(D–F)** ROC curves of the four binary classification tasks for the radiomic and DCNN models on the validation set.

Fig. 3. The t-SNE visualizations for the four classification tasks **(A–D)** of the standardized total collection of radiomic features (including age and gender) prior to feature selection, standardized radiomic features selected by LASSO, and DCNN features as described in Method. Every dot represents a patient. Red color represents the positive classes, whereas green color represents the negative classes.

Fig. 4. Visualizations of the correlations between the LASSO-selected radiomic features and the top 128 DCNN features for the four classification tasks **(A–D)**. Spearman's rank correlation coefficients (represented by ρ) were calculated between the radiomic and DCNN features. The pairs of radiomic and DCNN features were considered correlated and connected with yellow lines when the absolute values of ρ were > 0.5.

Fig. 5. Hierarchical diffuse glioma subtyping model. According to the 2016 WHO classification system, diffuse gliomas can be classified by a three-level subtyping model with four binary classification tasks: I, Grading; II, LGGs_IDH; III, LGGs_IDH_1p19q;

and IV, GBM_IDH. The molecular subtypes of diffuse gliomas are denoted in red text. Abbreviation: LGGs, lower-grade gliomas; GBM, glioblastoma.

Fig. 6. (A) Network structure of the developed 2.5D DCNN model for diffuse glioma subtyping. The input to the network is the tumor region of interest (ROI) cropped from three slices, each of which has three channels (T1w, T1CE, and T2w). The output is the probability of being classified as a positive sample. **(B)** Detailed network structure of the ResNet18 backbone. The layers with learnable parameters are represented as rectangular boxes. $n \times n$ indicates the kernel size, and $C \times W^2$ (in gray color) means that the current tensor size is $W \times W$ with C channels. The strides are one unless otherwise noted by the symbol '/2', which indicates that the stride is 2. The input to the network is a resized tumor ROI cropped from one slice; each ROI has three channels (T1w, T1CE, and T2w). The output is a 512-d feature vector. Abbreviations: MP = max pooling; fc = fully connected.

Supplementary Fig. 1. Performances of the radiomic **(A–D)** and DCNN **(E–H)** models for the four binary classification tasks on the five-fold cross-validation training set. ROC curves for each fold and the mean ROC curves.

Supplementary Fig. 2. Work flow of patient inclusion and exclusion.

Supplementary Fig. 3. The t-SNE visualizations using the validation data alone for the four classification tasks **(A–D)** of the standardized total collection of radiomic features (including age and gender) prior to feature selection; standardized radiomic features selected by LASSO; and DCNN features as described in Method. Every dot represents a patient. Red color represents the positive classes, whereas green color represents the negative classes. Despite much sparser data points, these visualizations generally reflect the same trends as in Figure 5.

Table 1. Patient characteristics in the training and validation sets. Please refer to Figure 1 for the molecular subtypes I–V.

Data set	Total number	Age		Sex		Molecular subtypes				
		Median	IQR [a]	Male	Female	I	II	III	IV	V
Training	780	47	21	445	335	138	116	191	60	275
Validation	236	47	19	132	104	47	33	47	19	90
P value	-	0.349 [b]		0.819 [c]		0.617 [c]				

[a] Interquartile range.
[b] Mann-Whitney U-test.
[c] Chi-square test.

Table 2. Performances of the radiomics and DCNN models on the entire training set. For the four classification tasks, the positive classes are

GBMs, LGGs with IDH mutation, 1p19q codeletion, and GBMs with IDH mutation, respectively.

Classification (subject number)	Model	AUC	Accuracy	Sensitivity	Specificity
Task I: Grading.	Radiomics	0.96	0.91	0.93	0.90
GBM vs. LGGs (335 vs. 445)	Radiomics ensemble	0.97	0.91	0.88	**0.92**
	DCNN	**0.99**	**0.94**	**0.97**	**0.92**
Task II: LGGs_IDH.	Radiomics	0.93	0.86	0.91	0.80
IDH mut vs. wt in LGGs (254 vs. 191)	Radiomics ensemble	0.93	0.84	0.82	0.87
	DCNN	**0.99**	**0.95**	**0.95**	**0.95**
Task III: LGGs_IDH_1p19q.	Radiomics	0.92	0.88	0.89	0.86
1p/19q codel vs. noncodel in IDH mut	Radiomics ensemble	0.92	0.87	0.88	0.85
LGGs (138 vs. 116)	DCNN	**1.00**	**0.95**	**0.99**	**0.91**
Task IV: GBM_IDH.	Radiomics	0.95	**0.92**	0.83	**0.94**
IDH mut vs. wt in GBMs (60 vs. 275)	Radiomics ensemble	0.95	0.89	0.85	0.90
	DCNN	**1.00**	0.90	**1.00**	0.88

Abbreviations: DCNN, deep convolutional neural network; GBM, glioblastoma; LGGs, lower-grade gliomas; vs., versus; mut, mutation; wt, wild type.

Numbers in bold font represent the best performance among different models.

Table 3. Performances of the established radiomics and DCNN models on the validation set.

Classification (Subject Number)	Model	AUC	Accuracy	Sensitivity	Specificity
Task I: Grading.	Radiomics	0.84	0.76	0.72	0.80
GBM vs. LGGs (335 vs. 445)	Radiomics ensemble	0.84	0.76	0.69	0.82
	DCNN	**0.89**	**0.83**	**0.81**	**0.84**
Task II: LGGs_IDH.	Radiomics	0.82	0.74	0.78	0.68
IDH mut vs. wt in LGGs (254 vs. 191)	Radiomics ensemble	0.82	0.71	0.68	0.77
	DCNN	**0.89**	**0.80**	**0.81**	**0.79**
Task III: LGGs_IDH_1p19q.	Radiomics	0.77	0.79	0.83	0.73
1p/19q codel vs. noncodel in IDH mut	Radiomics ensemble	0.78	0.75	0.79	0.70
LGGs (138 vs. 116)	DCNN	**0.85**	**0.83**	**0.85**	**0.79**
Task IV: GBM_IDH.	Radiomics	**0.68**	0.69	0.42	0.74
IDH mut vs. wt in GBMs (60 vs. 275)	Radiomics ensemble	0.67	0.64	**0.53**	0.67
	DCNN	0.66	**0.74**	0.47	**0.80**

Numbers in bold font represent the best performance among different models.

Table 4. Correlation analyses between the radiomic and DCNN features using the Spearman's rho (ρ).

Classification task	No. of Total feature	Correlated pairs /	Ratio of	LASSO weights [d]	DCNN feature weights [e]

	Setting	features [a]	pairs	features ($\|\rho\| > 0.5$) [b]	correlations (%) [c]	Min	Max	Min	Max	Threshold [f]
I: Grading	A	154:2,560	394,240	(6,982, 14, 1,353)	(1.77, 9.09, 52.85)	0.000	0.127	0.000	0.047	N.A.
	B	154:128	19712	(387, 9, 76)	(1.96, 5.84, 59.38)					0.041
II: LGG_IDH	A	87:2,560	222,720	(127, 16, 47)	(0.06, 18.39, 1.84)	0.000	0.071	0.000	0.047	N.A.
	B	87:128	11136	(6, 6, 2)	(0.05, 6.90, 1.56)					0.040
III: LGGs_	A	49:2,560	125,440	(860, 13, 359)	(0.69, 26.53, 14.02)	0.000	0.053	0.000	0.047	N.A.
IDH_1p19q	B	49:128	6272	(46, 9, 18)	(0.73, 18.37, 14.06)					0.041
IV: GBM-IDH	A	88:2,560	225,280	(61, 25, 21)	(0.03, 28,41, 0.82)	0.000	0.048	0.000	0.046	N.A.
	B	88:128	11264	(1, 1, 1)	(0.01, 1.14, 0.78)					0.041

[a] Setting A: the number of LASSO-selected radiomic features versus the total number of DCNN features, and B: the number of LASSO-selected radiomic features versus the top 128 DCNN features with largest absolute weights.

[b] Format (l, m, n): l is the total number of feature pairs that were correlated (identified by $|\rho| > 0.5$), m is the number of radiomic features correlated with DCNN features, and n is the number of DCNN features correlated with radiomic features.

[c] Format (r, r_r, r_c): $r =$ number of correlations / total number of feature pairs, $r_r =$ number of radiomic features correlated with DCNN features / total number of radiomic features, and $r_c =$ number of DCNN features corelated with radiomic features / total number of DCNN features used.

[d] Absolute values of nonzero LASSO weights.

[e] Absolute values of the DCNN feature weights extracted from the fully connected layers.

[f] The threshold value used to select the top 128 DCNN features.

Figure 1

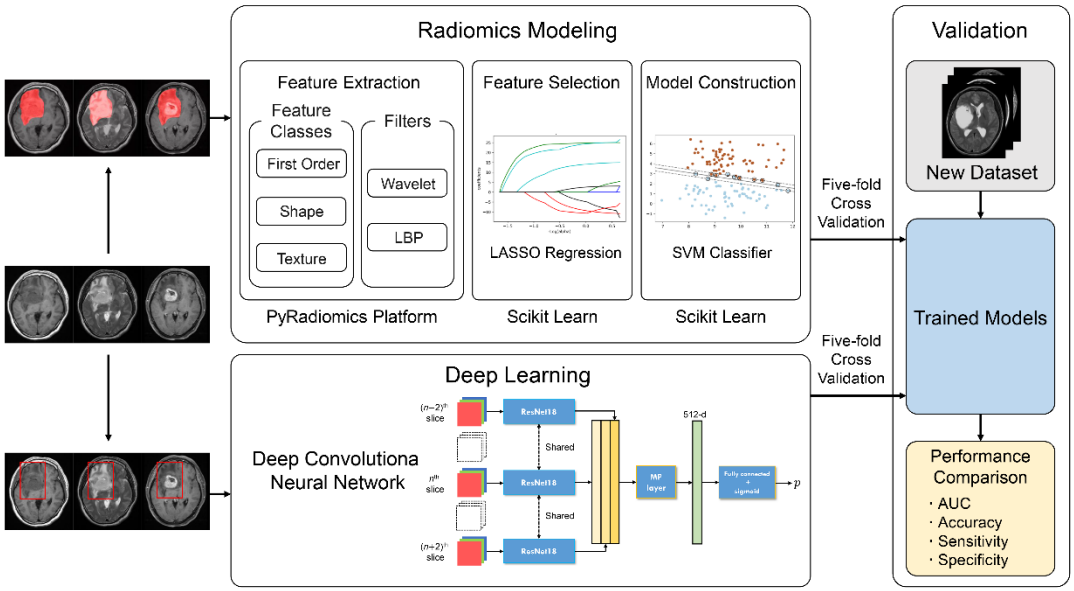


Figure 2

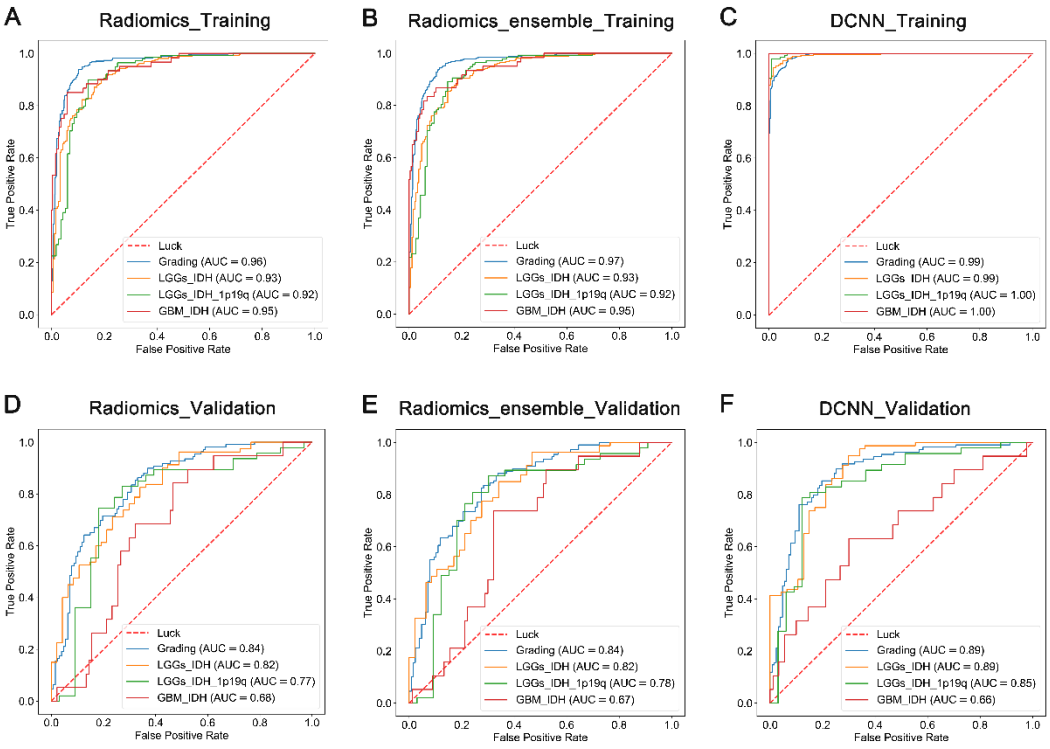


Figure 3

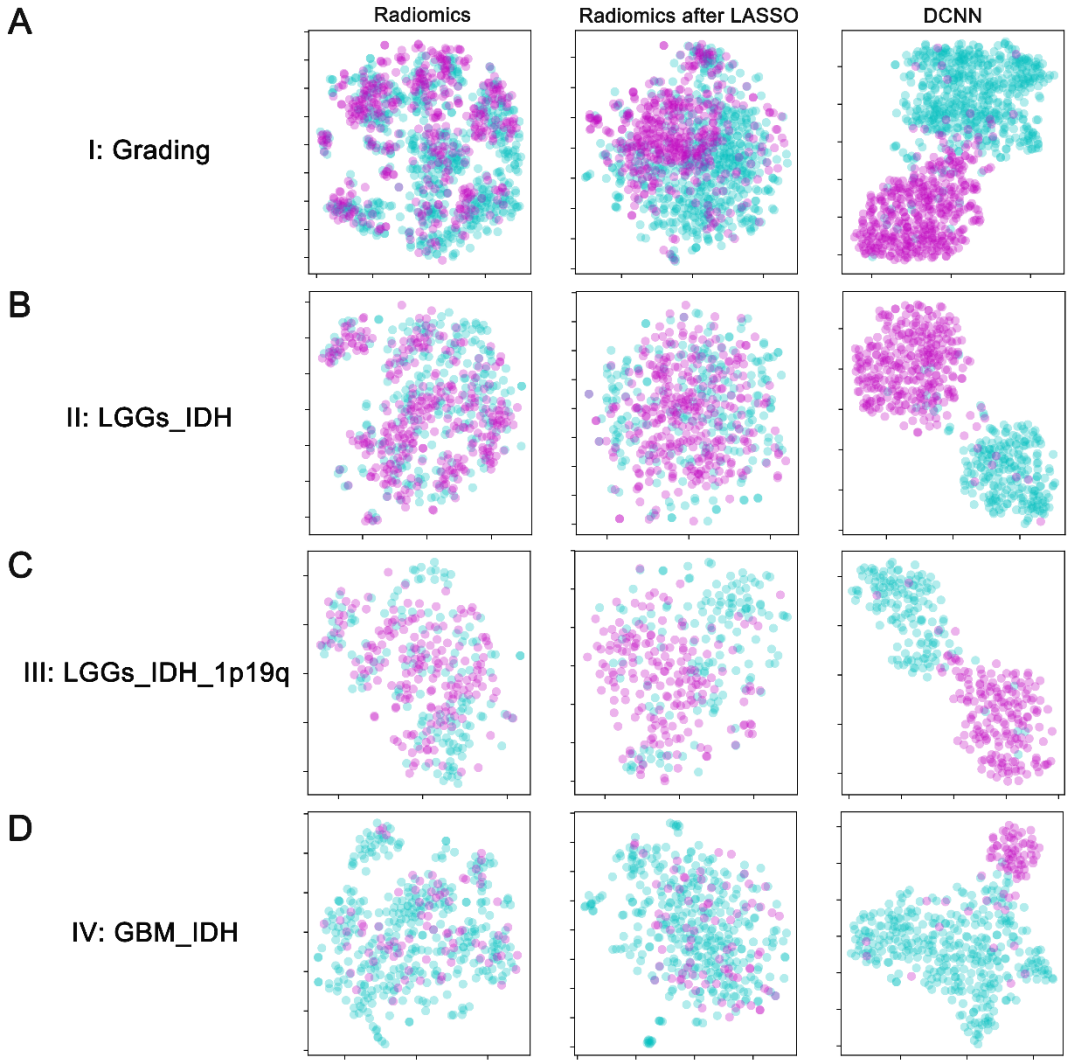

Radiomics
Radiomics after LASSO
DCNN
A
I: Grading
B
II: LGGs_IDH
C
III: LGGs_IDH_1p19q
D
IV: GBM_IDH

Figure 4

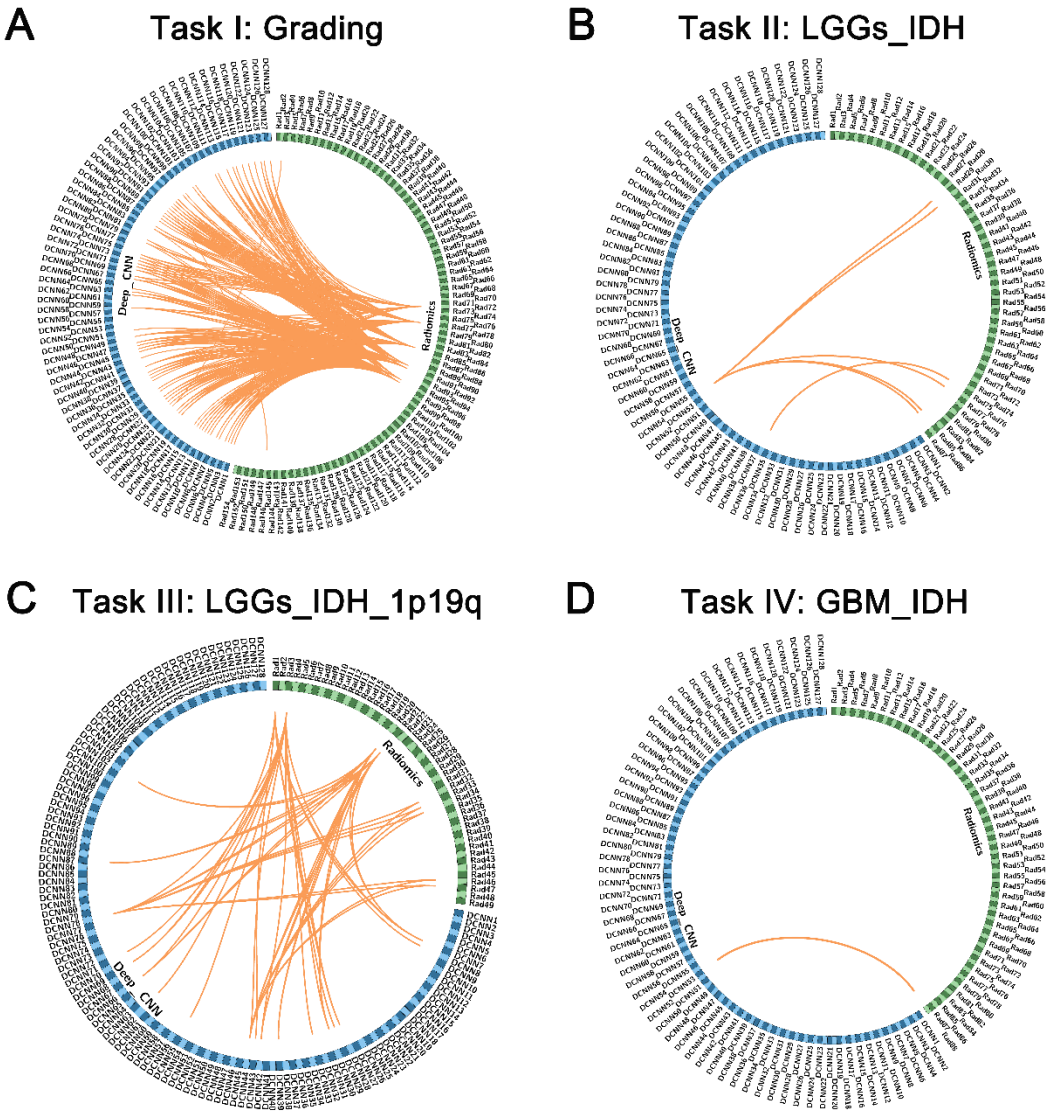

A
Task I: Grading
Deep_CNN
Radiomics
B
Task II: LGGs_IDH
Deep_CNN
Radiomics
C
Task III: LGGs_IDH_1p19q
Deep_CNN
Radiomics
D
Task IV: GBM_IDH
Deep_CNN
Radiomics

Figure 5

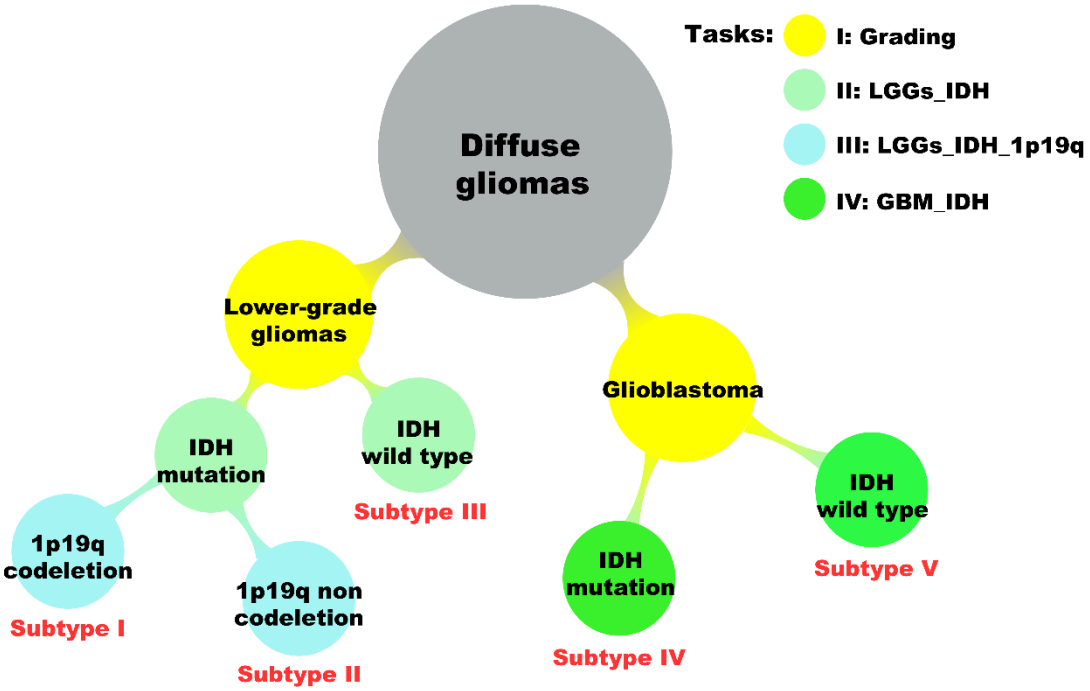
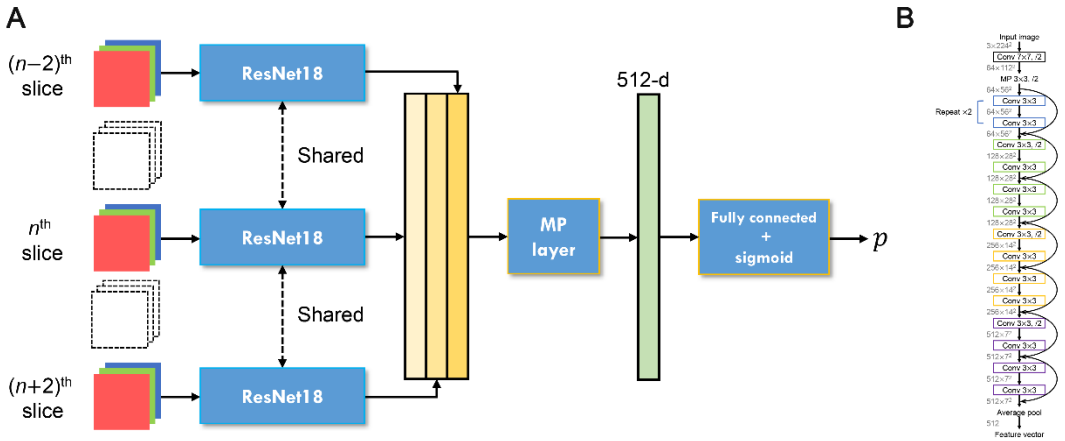


Figure 6



Supplementary Material

Radiomic feature extraction

For each case, 14 shape- and size-based features were calculated from the three-dimensional (3D) tumor mask. The 3D wavelet decomposition via directional low-pass and high-pass filtering was carried out on each original magnetic resonance (MR) sequence, yielding eight wavelet-transformed images. In addition, the local binary patterns (LBPs) were calculated in 3D, adding three LBP-transformed images. Then, 18 first-order and 75 textural features (including 24 gray level cooccurrence matrix features, 16 gray level run length matrix features, 16 gray level size zone matrix features, 14 gray level dependence matrix features, and five neighboring gray tone difference matrix features) were calculated on the original MR sequence, the eight wavelet-transformed images, and the three LBP-transformed images, resulting in 1,116 features, i.e., (18 + 75) features/image × (1 + 8 + 3) images. In total, 3,362 radiomic features were extracted from the three MR sequences (T1w, T1CE, and T2w), i.e., 1,116 features/sequence × 3 sequences + 14 shape and size features. For speed and memory consideration, the images were pre-cropped around the tumor regions prior to feature extraction. For detailed definition of these features, readers are referred to reference (*1*).

Deep convolutional neural network (DCNN) model structure

We developed a 2.5-dimensional (2.5D) DCNN model for the binary classification tasks of glioma subtyping (Fig. 6), on the basis of the residual blocks for deep learning (*2*). The residual blocks employ "skip connections" that bypass a few convolutional layers with an identity mapping of the input tensor. Consequently, the bypassed layers are encouraged to learn a residual that is added to the input tensor. This residual learning strategy achieved state-of-the-art performances on established benchmark data sets for natural image classification tasks. More specifically, we adapted a pretrained ResNet18 model (named for its 18 network layers with learnable parameters) as the backbone feature extractor in our DCNN model, by removing the average pooling and fully connected layers at the end. The same DCNN model was applied to the four binary classification tasks.

Considering the apparent anisotropy of the data used in this study, 2.5D input was used for the DCNN model (see the leftmost part of Fig. 6A) instead of 3D. Specifically, the slice with the maximum tumor area was first identified, assuming it was the n^{th} slice

of an input volume. Then, the $(n - 2)^{th}$, n^{th}, and $(n + 2)^{th}$ slices were extracted and input to the DCNN. We also experimented with three consecutive slices (i.e., $(n - 1)^{th}$, n^{th}, and $(n + 1)^{th}$ slices) instead of every other slices, but found this choice less stable. The three imaging modalities (T1w, T1CE, and T2w) were treated as three channels of a slice. A rectangular region of interest (ROI) that could cover the tumor areas in all the three slices was cropped out from each of these slices, and resized to 224×224 pixels. Next, the resized crops were input to three parallel, identical backbone networks with shared parameters, producing three feature vectors of 512 dimensions (512-d). Then, these three feature vectors were max-pooled (MP) across slices to become a single 512-d feature vector. The final feature vector was input to a fully connected (fc) layer and a sigmoid function, yielding p, the probability of being classified as a positive sample.

To mitigate class imbalance for a given binary classification task (if needed), the minority class with substantially fewer instances was over sampled by integer times to roughly match the number of instances in the majority class. Data augmentation was employed to counteract overfitting while training the DCNN models, including: (1) random selection of the central slice of the 2.5D input from the top three slices with largest tumor areas; (2) random translation within 10% and random scaling within [0.9, 1.1] with respect to the ROI to crop; (3) left/right mirroring of the cropped ROI with 50% probability; (4) random intensity scaling within [0.8, 1.2]; and (5) random rotation of the cropped ROI within [−10, 10] degrees. No class rebalance or data augmentation was used during validation or testing.

References

1. J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, H. Aerts, Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77, e104-e107 (2017).

2. K. He, Zhang, X., Ren, S. and Sun, J., Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778 (2016).

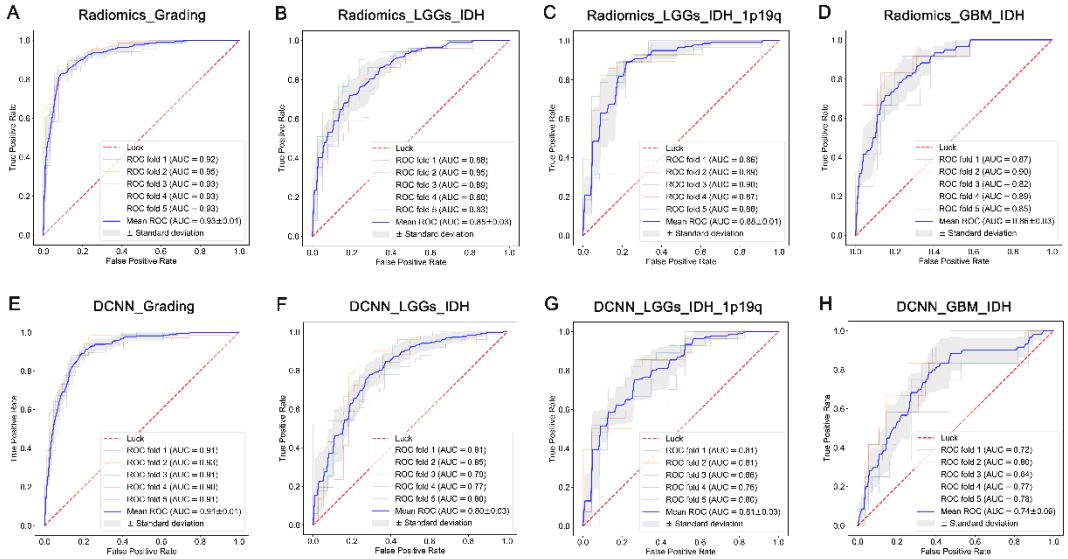
Supplementary Table 1. List of the radiomic features that were correlated with the DCNN features in Figure 4.

Classification task	Radiomic feature ID	Radiomic feature description	Correlated DCNN feature ID(s)
Task I: Grading	2	original_shape_Sphericity	24
	72	ce_wavelet-LHH_firstorder_Maximum	1; 2; 4; 7; 10; 11; 20; 21; 39; 44; 49; 58; 68; 71; 73; 75; 84; 89; 90; 91; 93; 94; 116
	77	ce_wavelet-HLL_firstorder_RootMeanSquared	1; 2; 4; 7; 9; 10; 11; 12; 13; 20; 21; 22; 25; 26; 27; 28; 31; 34; 35; 36; 37; 39; 44; 45; 46; 47; 49; 58; 62; 63; 67; 68; 70; 71; 72; 73; 74; 75; 79; 84; 85; 87; 88; 89; 90; 93; 94; 95; 96; 98; 102; 103; 111; 114; 116; 119; 121; 125; 126; 128
	85	ce_wavelet-HLH_firstorder_RootMeanSquared	1; 2; 4; 7; 9; 10; 11; 12; 13; 20; 21; 25; 26; 28; 31; 35; 36; 37; 39; 44; 47; 49; 63; 67; 68; 70; 71; 72; 73; 74; 75; 79; 84; 88; 89; 90; 93; 94; 95; 96; 102; 103; 111; 114; 116; 119; 121; 125; 126; 128
	91	ce_wavelet-HHH_firstorder_Energy	1; 2; 3; 4; 7; 9; 10; 11; 12; 13; 15; 20; 21; 22; 24; 25; 26; 27; 28; 30; 31; 35; 37; 39; 44; 45; 46; 47; 49; 51; 58; 62; 63; 67; 68; 70; 71; 72; 73; 75; 79; 80; 84; 88; 89; 90; 91; 92; 93; 94; 95; 96; 98; 102; 103; 110; 114; 116; 117; 121; 123; 125; 126; 128
	92	ce_wavelet-HHH_firstorder_TotalEnergy	1; 2; 3; 4; 7; 9; 10; 11; 12; 13; 15; 20; 21; 22; 24; 25; 26; 27; 28; 30; 31; 35; 37; 39; 44; 45; 46; 47; 49; 51; 58; 62; 63; 67; 68; 70; 71; 72; 73; 75; 79; 80; 84; 88; 89; 90; 91; 92; 93; 94; 95; 96; 98; 102; 103; 110; 114;

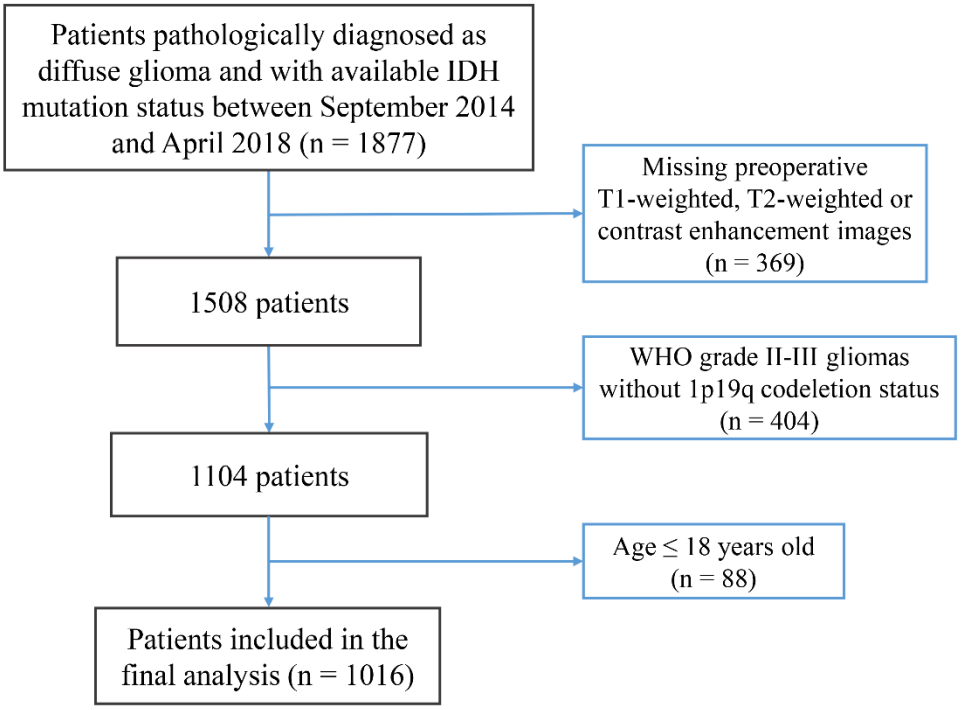
			116; 117; 121; 123; 125; 126; 128
	96	ce_wavelet-LLL_firstorder_Maximum	1; 2; 3; 4; 7; 9; 10; 11; 12; 13; 14; 20; 21; 22; 24; 25; 26; 27; 28; 31; 35; 36; 37; 39; 42; 44; 45; 46; 47; 49; 51; 58; 62; 63; 64; 67; 68; 70; 71; 72; 73; 74; 75; 79; 80; 84; 85; 88; 89; 90; 91; 92; 93; 94; 95; 96; 98; 99; 102; 103; 110; 111; 114; 116; 117; 119; 121; 125; 126; 128
	97	ce_wavelet-LLL_firstorder_Variance	1; 2; 4; 6; 7; 9; 10; 11; 12; 13; 20; 21; 22; 25; 26; 27; 28; 31; 35; 36; 37; 39; 44; 45; 47; 49; 58; 62; 63; 67; 68; 70; 71; 72; 73; 74; 75; 79; 84; 89; 90; 91; 93; 94; 95; 96; 98; 102; 103; 114; 116; 121; 125; 126
	146	t1_wavelet-LLL_glszm_ZoneVariance	91
Task II: LGGs_IDH	28	ce_wavelet-LLH_firstorder_Energy	54
	30	ce_wavelet-LLH_firstorder_TotalEnergy	54
	75	t1_wavelet-LLL_glszm_SizeZoneNonUniformity	36
	77	t1_wavelet-LLL_glszm_ZoneVariance	54
	85	t1_lbp-3D-k_gldm_LargeDependenceEmphasis	54
	86	t1_lbp-3D-k_gldm_LargeDependenceLowGrayLevelEmphasis	54
Task III: LGGs_IDH_1p19q	1	t2_original_firstorder_Skewness	7; 8; 28; 29; 43; 69
	3	t2_wavelet-LHL_glcm_Autocorrelation	7; 28; 30; 33; 42; 44; 55; 62; 66; 69; 71; 78
	20	ce_wavelet-HLL_gldm_HighGrayLevelEmphasis	7; 30; 39; 42; 44; 66; 78; 87; 102
	21	ce_wavelet-HLL_gldm_LowGrayLevelEmphasis	7; 30; 39; 42; 44; 66; 78; 87; 102
	32	t1_original_glcm_Autocorrelation	44; 78

	33	t1_original_glcm_JointAverage	44; 78
	34	t1_original_glcm_SumAverage	44; 78
	46	t1_wavelet-LLL_glcm_JointAverage	44; 78
	47	t1_wavelet-LLL_glcm_SumAverage	44; 78
Task IV: GBM_IDH	88	t1_lbp-3D-k_ngtdm_Busyness	53

Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3