Full title: Deep learning for non-invasive assessment of H3 K27M mutation status in diffuse midline gliomas using MR imaging

Abstract

BACKGROUND: Determination of H3 K27M mutation in diffuse midline glioma (DMG) is key for prognostic assessment and stratifying patient subgroups for clinical trials. MRI can noninvasively depict morphological and metabolic characteristics of H3 K27M mutant DMG.

PURPOSE: This study aimed to develop a deep learning (DL) approach to non-invasively predict H3 K27M mutation in DMG using T2-weighted images.

STUDY TYPE: Retrospective and prospective.

POPULATION: For diffuse midline brain gliomas, 341 patients from center-1 (27±19 years, 184 males), 42 patients from center-2 (33±19 years, 27 males) and 35 patients(37±18 years, 24males). For diffuse spinal cord gliomas, 133 patients from center-1 (30±15 years, 80 males).

FIELDSTRENGTH/SEQUENCE: 1.5T and 3T, T2-weighted turbo spin echo imaging.

ASSESSMENT: Conventional radiological features were independently reviewed by two neuroradiologists. H3 K27M status was determined by histopathological examination. The Dice coefficient was used to evaluate segmentation performance. Classification performance was evaluated using accuracy, sensitivity, specificity and area under the curve (AUC).

STATISTICAL TESTS: Pearson’s Chi-squared test, Fisher's exact test, two-sample Student’s t-test and Mann-Whitney U test. A two-sided P value < 0.05 was considered statistically significant.

RESULTS: In the testing cohort, Dice coefficients of tumor segmentation using DL were 0.87 for diffuse midline brain and 0.81 for spinal cord gliomas. In the internal prospective testing dataset, the predictive accuracies, sensitivities and specificities of H3 K27M mutation status were 92.1%, 98.2%, 82.9% in diffuse midline brain gliomas and 85.4%, 88.9%, 82.6% in spinal cord gliomas. Furthermore, this study showed that the performance generalizes to external institutions, with predictive accuracies of 85.7%-90.5%, sensitivities of 90.9%-96.0% and specificities of 82.4%-83.3%.

DATA CONCLUSION: In this study, an automatic DL framework was developed and validated for accurately predicting H3 K27M mutation using T2-weighted images, which could contribute to the noninvasive determination of H3 K27M status for clinical decision-making.

Evidence Level: 2

Technical Efficacy: Stage 2

Keywords: Diffuse midline glioma; H3 K27M mutation; Magnetic resonance imaging; Deep learning.

Introduction

Diffuse midline gliomas (DMGs) are a heterogeneous group of tumors involving the corpus callosum, thalamus, brainstem and spinal cord in both children and adults[1-5]. Genetic characterization identifies DMG patients with an unfavorable prognosis in those harboring a methionine mutation in histone H3 at lysine 27 (H3 K27M)[1,6-8]. Accurate identification of H3 K27M status contributes to diagnostic accuracy, prognostic assessment, stratification for clinical trials, and potential identification of individuals for targeted therapy (e.g. GD2-directed chimeric antigen receptor T cell therapy for H3 K27M-mutant diffuse intrinsic pontine gliomas and spinal cord DMGs)[1,7,9-11].

Currently, identification of H3 K27M-mutant status requires biopsy or surgery, which is time-consuming, expensive, and carries a risk of complications[7,12]. MRI provides a wealth of data to noninvasively depict tumor morphology and metabolic characteristics, which are associated with several genetic mutation in glioma, such as isocitrate dehydrogenase [IDH]-mutant, 1p/19q co-deletion, O^6-methylguanine-DNA-methytransferase (MGMT) methylation, and H3 K27M-mutant[13-19]. Therefore, accurate noninvasive and cost-effective determination of H3 K27M status using routine MRI has the potential to bypass the need for invasive biopsy[16,20], especially for those patients with contraindications.

Previous studies using conventional machine learning have confirmed the

ability of conventional MRI features to predict H3 K27M status of brainstem and spinal cord gliomas, with accuracies ranging from 60 to 85%[14,20-23]. Deep learning offers the ability to make this prediction without pre-engineered features, which has been particularly effective for other applications in glioma radiogenomics[24,25]. While a study had attempted to predict H3 K27M status of brainstem glioma, it was limited by a very small sample size (n = 55) without an external testing set or consideration of other locations of DMGs (e.g. thalamus or spinal cord), severely limiting the interpretation of the findings[26].

Indeed, the clinical translation of H3 K27M mutation prediction by MRI has been hampered by several factors, including (1) determination based on a small sample size (less than 100 cases), which may result in statistical bias and model overfitting; (2) lack of prospective and external validation sets, which may overestimate the generalizability of the predictive models; and (3) lack of an automatic pipeline from tumor segmentation to H3 K27M mutation prediction for DMGs (including the thalamus, brainstem, callosum and spinal cord).

This study aimed to establish an automatic DL pipeline, integrating segmentation and prediction, to accurately predict the H3 K27M status of DMGs (for both diffuse midline brain and spinal cord gliomas) based on a large multicenter dataset.

Materials and methods

Study Design And Participants

This study was in accordance with the Declaration of Helsinki and approved by the Animal and Human Ethics Committee of Beijing Tiantan Hospital, Capital Medical University. The external datasets were acquired upon approval by the local ethics committees Written informed consent was confirmed by each participant.

This was a multicenter study including retrospective and prospective data. MRI images retrospectively acquired during January 2018 and December 2019 (252 pathologically confirmed diffuse midline brain gliomas and 92 pathologically confirmed diffuse spinal cord gliomas) in Beijing Tiantan Hospital (center-1) were used to develop a DL framework. For retrospective data, this study had other inclusion criteria including (1) available axial T2WI for diffuse midline brain gliomas (thalamus, brainstem or callosum involved) or sagittal T2WI for intramedullary diffuse spinal cord gliomas; (2) available H3 K27M status; and (3) unifocal primary tumors prior to any clinical treatment. No further exclusion criteria for retrospective data. A prospectively acquired data of DMG was used for internal testing. For prospective data, the inclusion criteria including (1) clinically and radiologically suspected primary unifocal DMGs (diffuse midline brain or spinal cord gliomas) without any clinical treatment; and (2) patient confirmed to have biopsy or surgical resection. The exclusion criteria including

(1) pathologically confirmed non-gliomas; or (2) patients have no final biopsy or surgical resection. External testing was performed using two datasets from West China Hospital of Sichuan University (center-2: 42 diffuse midline brain gliomas) and Affiliated Jinling Hospital, Medical School of Nanjing University (center-3: 35 diffuse midline brain gliomas) (**Table 1, Table 2 and Figure 1**).

MR Acquisition

MR imaging was conducted on 1.5T or 3T MR scanners (including center-1: GE Signa HDxt and Discovery MR750, Siemens MAGNETOM Verio and Prisma, Philips Ingenia CX; center-2: GE Signa Excite, Siemens MAGNETOM Avanto, TrioTim and Skyra, Philips Achieva; center-3: GE Signa Discovery MR750, Siemens MAGNETOM TrioTim and Essenza). For diffuse midline brain glioma image, axial turbo-spin-echo (TSE) T2W images were acquired with the following protocol parameters: repetition time (TR)/echo time (TE) = 1800-12248 / 48-354 ms; flip angle (FA) = 90-160°; slice thickness = 1-6 mm; and matrix size = 256-640 × 256-640. The protocol parameters of the sagittal TSE T2W images in spinal cord glioma image were as follows: TR/TE = 1000-3060 / 48-130 ms; FA = 90-120°; slice thickness = 3-5 mm; and matrix size= 384-640 × 384-640 (**eTable 1 and eFigure 1**).

MRI Assessment

According to Visually Accessible REMBRANDT Images (VASARI) features and

the previous assessments of spinal cord glioma[27,28], conventional MRI features, including tumor location, edema, cystic/necrosis, hydrocephalus (diffuse midline brain gliomas), cavity (diffuse spinal cord gliomas) and contrast enhancement, were assessed independently by two neuroradiologists (Y. D, 14 years of experience in neuroradiology; M. W, 7 years of experience in neuroradiology), who were blinded to the genetic status (inter-rater agreement showed in **eTable 2**). This study presented and used conventional MRI features by Y. D.

Image Preprocessing

For brain images, this study used the following process[29]. T2W images were firstly skull-striped using 'BET' tool from FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Then the skull-striped T2W images were subjected to N4 bias correction using ANTs package (http://stnava.github.io/ANTs/). Then, the T2W images were cropped into the size of non-zero signal area. Finally, the signal intensities of each image were normalized by subtracting the mean and dividing by the image standard deviation. For spinal cord images, this study used the method reported in our previous study[30]. T2W images were cropped into the size of non-zero signal area and the signal intensities of each image were normalized by subtracting the mean and dividing the standard deviation.

DL Network for Tumor Segmentation

For DL tumor segmentation, the whole tumor region of the 252 diffuse midline brain gliomas and 92 diffuse spinal cord gliomas in the development dataset which were manually delineated by two neuroradiologists (L. Q and T. S, both with 4 years of experience in neuroradiology) independently using 3D Slicer (https://www.slicer.org/) and confirmed by Y. D, were used to train, validate, and test DL segmentation networks. For the diffuse midline brain tumor segmentation network, 151 tumors were used for training, 51 tumors for validation, and 50 tumors for testing. For the diffuse spinal cord tumor segmentation network, 55 tumors were used for training, 18 tumors for validation, and 19 tumors for testing. In both cases, separation of the training, validation, and testing sets was performed on the patient level. A 3D nnU-Net, was trained for tumor segmentation[31]. The DL tumor segmentation network architecture of nnU-Net is displayed in **Figure 2 and eDocument 1**. The Dice coefficient was used to evaluate segmentation performance.

DL Model for H3 K27M Mutation Prediction

For H3 K27M status prediction, two separate classification networks were developed for diffuse midline brain and spinal cord gliomas (**Figure 1C**). For diffuse midline brain gliomas, the prediction of H3 K27M mutation was based on the EfficientNet-B0 architecture, which is a network with parameter efficiency

while maintaining high predictive performance (**Figure 2**)[32]. For diffuse spinal cord gliomas, this study adopted a network with a simple architecture (**Figure 2 and eDocument 2**). To improve model robustness, the outputs of five DL networks derived from 5-fold cross-validation on the development dataset (retrospective dataset in center-1) were combined with the predictive output being determined with majority voting [$\geqslant$3/5], **Figure 1C** and e**Figure 2**). Additionally, the averaged predictive probability of five models followed by binarization was calculated for a comparison (**eTable 3**). Once the model was finalized, classification accuracy, sensitivity, specificity, and area under the receiver operator characteristic curve were used to evaluate predictive model performance on the independent internal and external testing datasets.

Ground Truth

H3 K27M status was determined by histopathological examination of biopsy or surgical samples, on which immunohistochemistry using H3 K27M mutation-specific antibody (1:500 dilution, Merck Millipore, Billerica, MA, USA) was performed.

Model Explanation

For a better understanding of the image features contributing to H3 K27M status prediction, Gradient-weighted Class Activation Mapping (Grad-CAM) was calculated and assessed independently by two neuroradiologists (L. Q and T.

S). Additionally, this study conducted sensitivity analysis for further interpretation of the model performance, including: (1) sub-group analyses of pediatric and adult cases were conducted to determinate whether the models perform differently in pediatric and adult patients, since evidences showed pediatric and adult DMGs have different pathogenies and prognosis[33,34]; (2) sub-group analysis on DMGs located in (a) corpus callosum, and (b) thalamus and brainstem separately, were conducted to determine a potential bias caused by tumor locations; (3) multiparametric MR analysis using available T2W-FLAIR and contrast-enhanced T1W images combined with T2W images to determine whether multiparametric MR modalities could improve the H3 K27M mutation predictive performance compared to those using only T2W images.

Statistical Analysis

SPSS (version 22, IBM, USA), MATLAB (version 2019a, MathWorks, USA), and Python (version 3.6) were used for statistical analyses. Categorical variables are displayed by ratios and tested using Pearson's Chi-squared test or Fisher's exact test. Continuous variables are displayed as the mean and standard deviation (SD) and tested using a two-sample Student's t-test if variable was normally distributed or Mann-Whitney U test if variable was not normally distributed. A two-sided $P<0.05$ was considered significant.

For comparison, conventional machine learning techniques were applied to explore the predictive ability of a combination of demographic (age, sex) and

manually radiologist-derived MRI features (tumor location, edema, cystic/necrosis, enhancement, hydrocephalus, cavity and whole tumor volume) for the H3 K27M mutation with identical development and testing datasets as those in DL pipelines. The conventional machine learning methods included lasso (least absolute shrinkage and selection operator) regression analysis, support vector machine, and multilayer perceptron.

Results

Demographics and clinical and conventional MRI characteristics

For prospective data, this study initially recruited 135 participants from January 2020 to September 2020. Five patients were excluded due to non-gliomas (n=3) or withdrawing clinical surgery (n=2). Finally, 130 patients were remained for internal testing in center-1 (**Table 1** and **2**, **Figure 1**).

A total of 474 DMGs, including retrospective data (252 diffuse midline brain gliomas and 92 diffuse spinal cord gliomas) and prospective data (89 diffuse midline brain gliomas and 41 diffuse spinal cord gliomas) in center-1, were included in this study. Two external datasets, including 42 diffuse midline brain gliomas in center-2 and 35 diffuse midline brain gliomas in center-3, were also included. In center-1, the patients with H3 K27M-mutant diffuse midline brain gliomas were younger than those with wild-type gliomas (aged 22±17 vs 40±17 years). Mutated tumors were predominately located in the thalamus (72/238 [30%] vs 16/103 [16%]) and brainstem (154/238 [65%] vs 21/103 [20%]) compared to wild-type cases. Lower frequencies of edema (47/238 [20%] vs 54/103 [52%]) and cystic/necrosis (122/238 [51%] vs 72/103 [70%]) and smaller whole tumor volumes (33±32 vs 75±67 ml) were observed in H3 K27M-mutant brain gliomas than in wild-type brain gliomas. Similar findings were also observed in center-2 and center-3. For diffuse spinal cord glioma, a lower presence ratio of cystic/necrosis (15/59 [25%] vs 45/74 [61%]), cavity (8/59

[14%] vs 32/74 [43%]) and enhancement (28/59 [47%] vs 56/74 [76%]) and a smaller whole tumor volume (8±5 vs 16±12 ml) were identified in H3 K27M-mutant cases than in wild-type glioma. Using lasso regression, this study confirmed the ability of these features to predict the H3 K27M mutation with accuracies ranging from 62.9% to 79.8% in prospective diffuse midline brain gliomas and 70.7% (95% CI-62.5% to 78.1%) in prospective diffuse spinal cord gliomas by majority voting (⩾3/5). Results using other conventional machine learning methods were found in **eTable 3**.

Tumor segmentation using DL

The Dice coefficients were 0.86 (95% CI-0.81 to 0.91) and 0.83 (95% CI-0.81 to 0.85) for diffuse midline brain and spinal cord gliomas, respectively, between the two expert raters. The Dice coefficients were 0.87 (95% CI-0.82 to 0.91) and 0.81 (95% CI-0.78 to 0.84) for diffuse midline brain and spinal cord gliomas, respectively, between DL and manual segmentations. No statistically significant difference was found between the Dice coefficients of tumor segmentation by the two raters and Dice coefficients of the DL and manual segmentations (For diffuse midline brain and spinal cord gliomas, the Dice coefficients were 0.86 vs. 0.87, P=0.782; 0.83 vs 0.81; P=0.172, respectively), demonstrating DL performance was within the range of inter-rater variability.

H3 K27M-mutation prediction using DL

For diffuse midline brain glioma, in the internal prospective testing dataset (center-1), the DL predictive network achieved an accuracy of 92.1% (95% CI-85.4% to 97.8%), sensitivity of 98.2% (95% CI-94.0% to 100%) and specificity of 82.9% (95% CI-69.4% to 94.3%). For external testing dataset 1 (center 2), the DL predictive network achieved an accuracy of 90.5% (95% CI-81.0% to 97.6%), sensitivity of 96.0% (95% CI-86.4% to 100%) and specificity of 82.4% (95% CI-62.5% to 100%). For the external testing dataset 2 (center-3), the DL predictive network achieved an accuracy of 85.7% (95% CI-74.3% to 97.1%), sensitivity of 90.9% (95% CI-84.6% to 100%) and specificity of 83.3% (95% CI-66.7% to 96.0%) (**Table 3**). Representative cases are shown in **Figure 3**. The classification performance of DL networks was superior to predictive models using demographic, clinical and conventional MRI features with accuracies of 79.8% (95% CI-76.1% to 84.5%) for center-1, 73.8% (95% CI-66.7% to 81.8%) for center-2, and 62.9% (95 CI-53.6% to 71.4%) for center-3.

For diffuse spinal cord gliomas, in the prospective testing dataset (center-1), the DL predictive network achieved an accuracy of 85.4% (95% CI-73.2% to 95.1%), sensitivity of 88.9% (95% CI-72.2% to 100%) and specificity of 82.6% (95% CI-65.4% to 96.0%) (**Table 3**). Representative cases are shown in **Figure 3**. The performance of DL networks was higher than predictive models using demographic, clinical, and conventional MRI features, which achieved an accuracy of 70.7% (95% CI-62.5% to 78.1%) (*P*=.094).

Model explanation

For diffuse midline brain glioma, the Grad-CAM (**Figure 4** and **eTable 4**) showed that the main activation areas were tumor core and peritumoral areas. Sub-group analyses (**eTable 5**) showed that the predictive performance of H3 K27M diffuse midline brain gliomas were consistent with the main findings (accuracies of 85.7%-92.1%) with accuracies of 81.5%-100% in both pediatric and adult groups. The model achieved predictive accuracies of 90.0%-100% for DMGs located in corpus callosum and 81.8%-93.2% for DGMs located in thalamus and brainstem (**eTable 5**), consistent with the main findings. Multiparametric MR images analysis showed that the predictive performances (accuracies of 51.7%-81.7%) using a combination of T2W, T2W-FLAIR and contrast-enhanced T1W images tended to be inferior to those (accuracies of 85.7%-92.1%) using only T2W images, indicating limited or deleterious contribution of T2W-FLAIR and contrast-enhanced T1W images to the determination of H3 K27M-mutant diffuse midline brain gliomas (**eTable 6**).

For diffuse spinal cord glioma, the Grad-CAM (**Figure 4** and **eTable 4**) showed that the main activation areas involved the entire tumor and peritumoral areas. Sub-group analyses (**eTable 5**) showed that the predictive performance of H3 K27M-mutant diffuse spinal cord gliomas tend to decrease with an accuracy of 78.6% in pediatric gliomas compared to the main finding (accuracy of 85.4%).

Multimodal MR images analysis showed that the predictive performance (accuracy of 82.5%) using a combination of T2W and contrast-enhanced T1W images was comparable to that (accuracy of 85.4%) using only T2W images (**eTable 6**).

Discussion

This study developed an automatic DL pipeline for H3 K27M status prediction in DMGs using multicenter datasets. Among the key findings, this study found high segmentation performances of brain and spinal cord tumors for the DL model within the range of inter-rater variability. Secondly, this study showed high predictive accuracies for mutation status in the internal prospective and external testing sets, exceeding the performance of conventional machine learning models using demographics and radiologist-derived MR features. Importantly, the DL pipeline only utilizes T2W MR imaging, which is acquired under standard of care, without the need for advanced MR sequences or other modalities.

The investigations of noninvasive radiological findings (radiomics and deep learning) to predict H3 K27M status are growing[20,21,23,26]. The first major obstacle is tumor segmentation prior to genotype prediction. A large number of deep learning networks have been developed for brain whole tumor segmentation with overall good performance (Dice coefficients ranging from 0.80 to 0.91) using multiple MRI sequences, including T2WI, T1WI, contrast-enhanced T1WI and T2WI-FLAIR[35,36], but few studies have focused specifically on diffuse midline brain or spinal cord glioma segmentation due to their relative rarity[37]. In our study, the nnU-Net network, which is a state-of-the-art architecture for various segmentation tasks[31], was used for diffuse midline brain

and spinal cord tumor segmentation. The performance for whole tumor segmentation was excellent for diffuse midline brain tumors, comparable to inter-rater variability of manual segmentation and previous brain tumor segmentations (Dice coefficients ranging from 0.80 to 0.91). The segmentation of spinal cord tumors was comparable to a previous segmentation task of spinal cord tumors (Dice coefficients ranging from 0.77 to 0.80)[37]. Accurate tumor segmentation contributes to H3 K27M status prediction accuracy, since tumor morphology and location were potential predictors for H3 K27M mutation[1].

For diffuse midline brain and spinal cord gliomas, two separate DL pipelines were combined for H3 K27M mutation prediction. Given that the number of patients with H3 K27M-mutant gliomas was significantly larger than that of wild-type patients (H3 K27M-mutation percentage achieved 69.8% in center-1), this study integrated the final output from 5-fold cross-validations derived from the development dataset (majority voting [$\geqslant$3/5] as the final prediction, which was comparable to using averaged predictive probability) to improve the network robustness. The performance of the H3 K27M status prediction in diffuse midline brain gliomas achieved an accuracy of 92.1% in the internal prospective testing dataset, and the model performance was maintained within external testing datasets (accuracy of 85.7%-90.5%), subgroup analyses in pediatric and adult groups (accuracy of 81.5%-100%), and DMGs located in corpus callosum, thalamus and brainstem (accuracy of 81.8%-100%), indicating

robustness across various sites, scanners, field strength, age groups and tumor locations, and demonstrating its generalizability for clinical translation. The performance of DL model using multimodal MR images (T2W-FLAIR and contrast-enhanced T1W images) was not superior to the current model using only T2W images, indicating that the features extracted from T2W image were efficient to predicting H3 K27M-mutant DMGs. Based on demographics (e.g. age) and conventional MRI features of tumors (e.g. tumor location), an accuracy of approximately 70% for predicting H3 K27M status could be reached using conventional machine learning algorithms, including lasso regression, support vector machine and multilayer perceptron. The DL algorithms had a 15-20% higher accuracy than the routine machine learning methods and used only T2W images. The potential explanation may be that the T2WI information extracted by DL can capture the tumor features (e.g. features of tumor core and peritumoral areas) and morphology, providing more accurate image interpretation for H3 K27M-mutant gliomas than conventional clinical and MRI features.

A study has been conducted on the prediction of H3 K27M alterations in diffuse spinal cord gliomas, with a prediction accuracy less than 65%[22]. In the current study, an accuracy of 85.4% was reached, similar to that (accuracy of 82.5%) using a combination of T2W and contrast-enhanced T1W images, showing a tendency toward higher accuracy compared to predictive models using

demographics and conventional MRI features with an accuracy of 70.7% (*P*=.094, which is not statistically significant due to the small testing sample size [n=41]). The current findings with regard to diffuse spinal cord gliomas are promising for rapid and accurate clinical diagnosis and personalized treatment intervention but warrant further validation with more samples.

LIMITATIONS

First, the current molecular predictive task focused on the H3 K27M-mutant DMGs. The latest definition of “diffuse midline gliomas, H3 K27-altered” involves additional molecular alterations (e.g. EZHIP), which demonstrated similar responses to treatment and clinical outcomes as those harboring H3 K27M mutation[38,39]. These new molecular subtypes could be considered in future studies to transfer and expand the current DL models towards prediction of “diffuse midline glioma, H3 K27-altered”. Second, while the utilization of multiparametric structural MRI (combination of T2W, T2W-FLAIR and contrast-enhanced T1W images) did not improve the current model (data not shown), other advanced MR sequences (e.g. diffusion imaging and perfusion imaging) might contribute to an improvement in H3 K27M status prediction. Additionally, heterogeneous MR scan parameters may limit the deep learning algorithm accuracy, which should be considered in the future work. Third, even though the segmentation and prediction networks performed well in the current study, more efforts on the network design (e.g. joining the segmentation and prediction

task into one network architecture) could be explored to improve performance and efficiency of the models. Last, prospective testing was only performed in histopathologically confirmed DMGs in center 1 and the sample size of diffuse spinal cord gliomas was relatively small. Future studies could evaluate the performance of the established framework in clinical practice by integrating a prior differentiation task of determining diffuse midline glioma from other tumors.

CONCLUSION

An automatic DL framework was developed and externally validated for accurately segmenting DMGs in both the brain and spinal cord and subsequently predicting H3 K27M mutation, which can contribute to the noninvasive identification of H3 K27M status to improve patient management.

References

1. LOUIS D N, PERRY A, REIFENBERGER G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 2016;**131**(6):803-820.
2. KHUONG-QUANG D A, BUCZKOWICZ P, RAKOPOULOS P, et al. K27M mutation in histone H3.3 defines clinically and biologically distinct subgroups of pediatric diffuse intrinsic pontine gliomas. Acta Neuropathol 2012;**124**(3):439-447.
3. WU G, BRONISCER A, MCEACHRON T A, et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. Nat Genet 2012;**44**(3):251-253.
4. ENOMOTO T, AOKI M, HAMASAKI M, et al. Midline Glioma in Adults: Clinicopathological, Genetic, and Epigenetic Analysis. Neurol Med Chir (Tokyo) 2020;**60**(3):136-146.
5. WANG L, LI Z, ZHANG M, et al. H3 K27M-mutant diffuse midline gliomas in different anatomical locations. Hum Pathol 2018;**78**:89-96.
6. JONES C, KARAJANNIS M A, JONES D T W, et al. Pediatric high-grade glioma: biologically and clinically in need of new thinking. Neuro Oncol 2017;**19**(2):153-161.
7. WIERZBICKI K, RAVI K, FRANSON A, et al. Targeting and Therapeutic Monitoring of H3K27M-Mutant Glioma. Curr Oncol Rep 2020;**22**(2):19.
8. LOUIS D N, GIANNINI C, CAPPER D, et al. cIMPACT-NOW update 2:

diagnostic clarifications for diffuse midline glioma, H3 K27M-mutant and diffuse astrocytoma/anaplastic astrocytoma, IDH-mutant. Acta Neuropathol 2018;**135**(4):639-642.

9. FILBIN M G, TIROSH I, HOVESTADT V, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science 2018;**360**(6386):331-335.
10. CASTEL D, PHILIPPE C, CALMON R, et al. Histone H3F3A and HIST1H3B K27M mutations define two subgroups of diffuse intrinsic pontine gliomas with different prognosis and phenotypes. Acta Neuropathol 2015;**130**(6):815-827.
11. MAJZNER R G, RAMAKRISHNA S, YEOM K W, et al. GD2-CAR T cell therapy for H3K27M-mutated diffuse midline gliomas. Nature 2022;**603**(7903):934-941.
12. PINCUS D W, RICHTER E O, YACHNIS A T, et al. Brainstem stereotactic biopsy sampling in children. J Neurosurg 2006;**104**(2 Suppl):108-114.
13. REES J H. Diagnosis and treatment in neuro-oncology: an oncological perspective. Br J Radiol 2011;**84** Spec No 2:S82-S89.
14. PICCARDO A, TORTORA D, MASCELLI S, et al. Advanced MR imaging and (18)F-DOPA PET characteristics of H3K27M-mutant and wild-type pediatric diffuse midline gliomas. Eur J Nucl Med Mol Imaging 2019;**46**(8):1685-1694.
15. CALMON R, DANGOULOFF-ROS V, VARLET P, et al. Radiogenomics

of diffuse intrinsic pontine gliomas (DIPGs): correlation of histological and biological characteristics with multimodal MRI features. Eur Radiol 2021;**31**(12):8913-8924.

16. LEACH J L, ROEBKER J, SCHAFER A, et al. MR imaging features of diffuse intrinsic pontine glioma and relationship to overall survival: report from the International DIPG Registry. Neuro Oncol 2020;**22**(11):1647-1657.

17. PATEL S H, BATCHALA P P, MRACHEK E K S, et al. MRI and CT Identify Isocitrate Dehydrogenase (IDH)-Mutant Lower-Grade Gliomas Misclassified to 1p/19q Codeletion Status with Fluorescence in Situ Hybridization. Radiology 2020;**294**(1):160-167.

18. PATEL S H, POISSON L M, BRAT D J, et al. T2-FLAIR Mismatch, an Imaging Biomarker for IDH and 1p/19q Status in Lower-grade Gliomas: A TCGA/TCIA Project. Clin Cancer Res 2017;**23**(20):6078-6085.

19. SUH C H, KIM H S, JUNG S C, et al. Clinically Relevant Imaging Features for MGMT Promoter Methylation in Multiple Glioblastoma Studies: A Systematic Review and Meta-Analysis. AJNR Am J Neuroradiol 2018;**39**(8):1439-1445.

20. PAN C C, LIU J, TANG J, et al. A machine learning-based prediction model of H3K27M mutations in brainstem gliomas using conventional MRI and clinical features. Radiother Oncol 2019;**130**:172-179.

21. KANDEMIRLI S G, KOCAK B, NAGANAWA S, et al. Machine Learning-

Based Multiparametric Magnetic Resonance Imaging Radiomics for Prediction of H3K27M Mutation in Midline Gliomas. World Neurosurg 2021;**151**:e78-e85.

22. JUNG J S, CHOI Y S, AHN S S, et al. Differentiation between spinal cord diffuse midline glioma with histone H3 K27M mutation and wild type: comparative magnetic resonance imaging. Neuroradiology 2019; **61**(3):313-322.

23. SU X, CHEN N, SUN H, et al. Automated machine learning based on radiomics features predicts H3 K27M mutation in midline gliomas of the brain. Neuro Oncol 2020;**22**(3):393-401.

24. CHANG K, BAI H X, ZHOU H, et al. Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging. Clin Cancer Res 2018;**24**(5):1073-1081.

25. CHANG P, GRINBAND J, WEINBERG B D, et al. Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas. AJNR Am J Neuroradiol 2018;**39**(7):1201-1207.

26. LIU J, CHEN F, PAN C, et al. A Cascaded Deep Convolutional Neural Network for Joint Segmentation and Genotype Prediction of Brainstem Gliomas. IEEE Trans Biomed Eng 2018;**65**(9):1943-1952.

27. TUSTISON N J, COOK P A, KLEIN A, et al. Large-scale Evaluation of ANTs and FreeSurfer Cortical Thickness Measurements. Neuroimage, 2014, **99**:166-179.

28. Wiki for the VASARI feature set The National Cancer Institute Web site. Available at https://wikicancerimagingarchivenet/display/Public/VASARI+Research+Project Updated March 20, 2020

29. KIM DH, KIm JH, CHOI SH, et al. Differentiation between intramedullary spinal ependymoma and astrocytoma: comparative MRI analysis. Clin Radiol. 2014;**69**(1):29-35.

30 ZHUO Z, ZHANG J, DUAN Y, et al. Automated Classification of Intramedullary Spinal Cord Tumors and Inflammatory Demyelinating Lesions Using Deep Learning. Radiol Artif Intell, 2022, **4**(6):e210292.

31. ISENSEE F, JAEGER P F, KOHL S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods, 2021;**18**(2):203-211.

32. V.LE M T Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning 2019;**97**:6105–6114.

33. JIANG H, YANG K, REN X, et al. Diffuse midline glioma with H3 K27M mutation: a comparison integrating the clinical, radiological, and molecular features between adult and pediatric patients. Neuro Oncol 2020;**22**(5):e1-e9.

34. SCHULTE J D, BUERKI R A, LAPOINTE S, et al. Clinical, radiologic, and genetic characteristics of histone H3 K27M-mutant diffuse midline

gliomas in adults. Neurooncol Adv 2020;**2**(1):vdaa142.

35. ANGULAKSHMI M, DEEPA M. A Review on Deep Learning Architecture and Methods for MRI Brain Tumour Segmentation. Curr Med Imaging 2021;**17**(6):695-706.

36. MAGADZA T, VIRIRI S. Deep Learning for Brain Tumor Segmentation: A Survey of State-of-the-Art. J Imaging 2021;**7**(2):19.

37. ANDREANNE LEMAY C G, ZHIZHENG ZHUO, JIE ZHANG, et al. Multiclass Spinal Cord Tumor Segmentation on MRI with Deep Learning. Neuroimage Clin 2021;**31**:102766.

38. SIEVERS P, SILL M, SCHRIMPF D, et al. A subset of pediatric-type thalamic gliomas share a distinct DNA methylation profile, H3K27me3 loss and frequent alteration of EGFR. Neuro Oncol 2021;**23**(1):34-43.

39. CASTEL D, KERGROHEN T, TAUZIEDE-ESPARIAT A, et al. Histone H3 wild-type DIPG/DMG overexpressing EZHIP extend the spectrum diffuse midline gliomas with PRC2 inhibition beyond H3-K27M mutation. Acta Neuropathol 2020;**139**(6):1109-1113.

Tables

Table 1. Demographics and conventional MRI features of diffuse midline brain gliomas.

	Center-1											Center-2			Center-3		
	Training, Validation and internal prospective testing		*P* value	Training		*P* value	Validation		*P* value	internal prospective testing		*P* value			*P* value	external testing		*P* value
	H3 K27M-mutant (n=238)	Wild-type (n=103)		H3 K27M-mutant (n=169)	Wild-type (n=53)		H3 K27M-mutant (n=15)	Wild-type (n=15)		H3 K27M-mutant (n=54)	Wild-type (n=35)		H3 K27M-mutant (n=25)	Wild-type (n=17)		H3 K27M-mutant (n=11)	Wild-type (n=24)	
Demographics																		
Age (mean±SD, year)	22±17	40±17	<0.001	22±17	41±15	<0.001	19±14	36.1±19.3	0.019	22±16	39±19	<0.001	27±18	40±18	0.030	31±15	40±18	0.177
(Female/Male)	105/133	52/51	0.279	71/98	30/23	0.063	6/9	8/7	0.464	28/26	14/21	0.274	9/16	6/11	0.963	3/8	8/16	>0.999
Location																		
Corpus callosum(n, %)	12 (5%)	66 (64%)	<0.001	5 (3.0%)	31 (58%)	<0.001	3 (20%)	9 (60%)	0.025	4 (7.4%)	26 (74%)	<0.001	2 (8%)	6 (35%)	0.045	3 (27%)	10 (42%)	0.478
Thalamus(n, %)	72 (30%)	16 (16%)	0.004	49 (29%)	8 (15%)	0.043	4 (27%)	4 (27%)	>0.999	19 (35%)	4 (11%)	0.012	8 (32%)	5 (29%)	0.859	4 (36%)	4 (17%)	0.226
Brainstem(n, %)	154 (65%)	21 (20%)	<0.001	115 (68%)	14 (26%)	<0.001	8 (53%)	2 (13%)	0.02	31 (57%)	5 (14%)	<0.001	15 (60%)	6 (35%)	0.116	4 (36%)	10 (42%)	>0.999
MRI presentation																		
Edema(n, %)	47 (20%)	54 (52%)	<0.001	32 (19%)	26 (49%)	<0.001	4 (27%)	7 (47%)	0.256	11 (20%)	21 (60%)	<0.001	8 (32%)	10 (59%)	0.085	3 (27%)	12 (50%)	0.281
Cystic/necrosis(n, %)	122 (51%)	72 (70%)	0.001	83 (49%)	40 (75%)	<0.001	9 (60%)	11 (73%)	0.439	30 (56%)	21 (60%)	0.679	13 (52%)	7 (41%)	0.491	7 (64%)	15 (63%)	>0.999
Enhancement(n, %)	151 (63%)	69 (67%)	0.53	110 (65%)	38 (72%)	0.373	9 (60%)	10 (67%)	0.705	32 (59%)	21 (60%)	0.945	14 (56%)	13 (76%)	0.174	7 (64%)	15 (63%)	>0.999
Hydrocephalus(n, %)	66 (28%)	24 (23%)	0.394	45 (27%)	10 (19%)	0.254	4 (27%)	1 (6.7%)	0.33	17 (31%)	13 (37%)	0.581	8 (32%)	1 (6%)	0.060	4 (36%)	3 (13%)	0.171
Whole tumor volume (mean±SD, ml)	33±32	75±67	<0.001	32±32	63±63	0.033	34±28	65±51	0.067	35±32	97±74	<0.001	32±39	66±81	0.446	20±25	52±42	0.044

Note: Whole tumor volume was calculated using manual (development dataset) or deep learning (testing datasets) segmentations. n, number; SD,

standard deviation.

Table 2. Demographics and conventional MRI features of diffuse spinal cord gliomas.

	Training, Validation and internal prospective testing		*P* value	Training		*P* value	Validation		*P* value	internal prospective testing		*P* value
	H3 K27M-mutant (n=59)	Wild-type (n=74)		H3 K27M-mutant (n=37)	Wild-type (n=46)		H3 K27M-mutant (n=4)	Wild-type (n=5)		H3 K27M-mutant (n=18)	Wild-type (n=23)	
Demographics												
Age (mean±SD, year)	32±12	29±16	0.152	32±13	29±16	0.248	42±8	25±10	0.444	30±12	30±18	0.990
Sex (Female/Male)	30/29	23/51	0.021	17/20	17/29	0.408	2/2	0/5	0.167	11/7	6/17	0.024
Location												
Cervical (n, %)	17 (29%)	31 (42%)	0.119	13 (35%)	19 (41%)	0.566	1 (25%)	3 (60%)	0.524	3 (17%)	9 (39%)	0.117
Cervical-thoracic (n, %)	9 (15%)	10 (14%)	0.776	6 (16%)	6 (13%)	0.683	1 (25%)	1 (20%)	>0.999	2 (11%)	3 (13%)	>0.999
Thoracic (n, %)	22 (37%)	26 (35%)	0.797	12 (32%)	15 (33%)	0.986	1 (25%)	1 (20%)	>0.999	9 (50%)	10 (43%)	0.678
Thoracic-lumbar (n, %)	11 (19%)	7 (9%)	0.124	6 (16%)	6 (13%)	0.683	1 (25%)	0 (0%)	0.444	4 (22%)	1 (4.3%)	0.150
MRI presentation												
Edema (n, %)	45 (76%)	55 (74%)	0.796	29 (78%)	33 (72%)	0.489	2 (50%)	5 (100%)	0.167	14 (78%)	17 (74%)	>0.999
Cystic/necrosis (n, %)	15 (25%)	45 (61%)	<0.001	12 (32%)	27 (59%)	0.017	0 (0%)	5 (100%)	0.008	3 (17%)	13 (57%)	0.009
Enhancement (n, %)	28 (47%)	56 (76%)	<0.001	20 (54%)	36 (78%)	0.019	2 (50%)	4 (80%)	0.524	6 (33%)	16 (70%)	0.021
Cavity (n, %)	8 (14%)	32 (43%)	<0.001	5 (14%)	23 (50%)	<0.001	1 (25%)	2 (40%)	>0.999	2 (11%)	7 (30%)	0.254
Whole tumor volume (mean±SD, ml)	8±5	16±12	<0.001	9±6	16±12	0.006	6±2	16±11	>0.999	7±5	15±12	0.056

Note: spinal cord gliomas were only acquired at Center-1. Whole tumor volume was calculated using manual (development dataset) or deep learning (testing datasets) segmentations. n, number; SD, standard deviation.

Table 3. The performance of lasso regression models for H3 K27M mutation prediction using demographics and conventional MRI features and DL models using conventional T2W images by majority voting (3/5).

		Accuracy (%, [95% CI])	Sensitivity (%, [95% CI])	Specificity (%, [95% CI])
Lasso regression using demographics and conventional MRI features	**Diffuse midline brain glioma**[a]			
	Center-1 testing dataset	79.8 (76.1-84.5)	79.6 (74.4-85.4)	80.0 (74.1-88.0)
	Center-2 testing dataset	73.8 (66.7-81.8)	84.0 (77.8-94.4)	58.8 (46.2-72.7)
	Center-3 testing dataset	62.9 (53.6-71.4)	54.6 (37.5-71.4)	66.7 (57.9-77.8)
	Diffuse spinal cord glioma[b]			
	Center-1 testing dataset	70.7 (62.5-78.1)	83.3 (76.9-92.9)	60.9 (50.0-70.6)
DL using T2WI	**Diffuse midline brain glioma**			
	Center-1 testing dataset	92.1 (85.4-97.8)	98.2 (94.0-100)	82.9 (69.4-94.3)
	Center-2 testing dataset	90.5 (81.0-97.6)	96.0 (86.4-100)	82.4 (62.5-100)
	Center-3 testing dataset	85.7 (74.3-97.1)	90.9 (84.6-100)	83.3 (66.7-96.0)
	Diffuse spinal cord			

glioma				
Center-1 testing dataset	85.4 (73.2-95.1)	88.9 (72.2-100)	82.6 (65.4-96.0)	

Note: CI, confidence interval; DL, deep learning; AUC, area under the curve; Lasso, Least absolute shrinkage and selection operator.

CI calculation used a bootstrap method in testing datasets.

[a] The significant contributing features for H3 K27M mutation prediction in diffuse midline brain glioma were tumor location at the brainstem and thalamus and younger age;

[b] The predominant contributing features for H3 K27M mutation prediction in diffuse spinal cord glioma were smaller whole tumor volume and absence of cavity.

Figure Legends

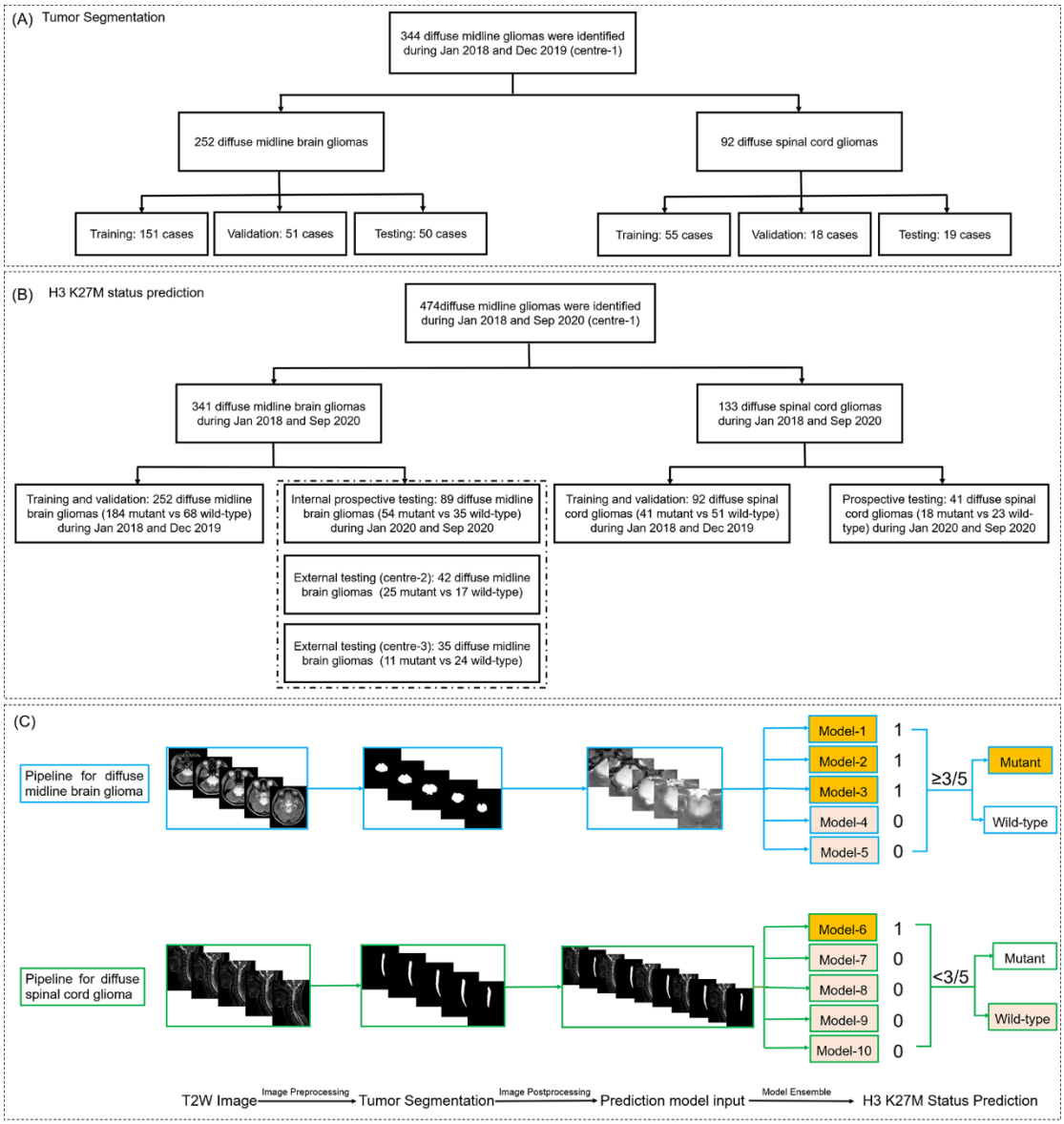


Figure 1. A flow chart of the included patients with diffuse midline glioma for (**A**) tumor segmentation and (**B**) H3 K27M status prediction. (**C**) Automatic framework integrating two DL pipelines for diffuse midline brain and spinal cord gliomas.

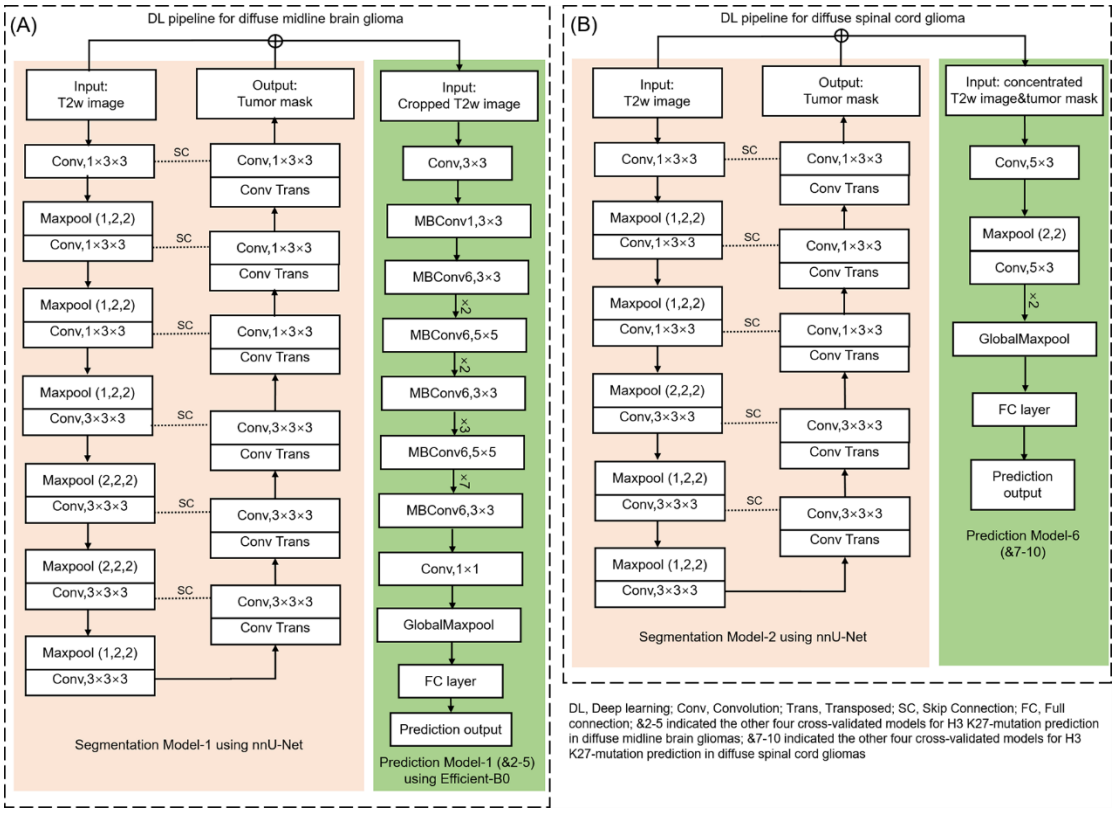


Figure 2. Automatic framework integrating two deep learning pipelines for diffuse midline brain and spinal cord gliomas based on nnU-Net, Efficient-B0 and an architecture of simple three layers networks.

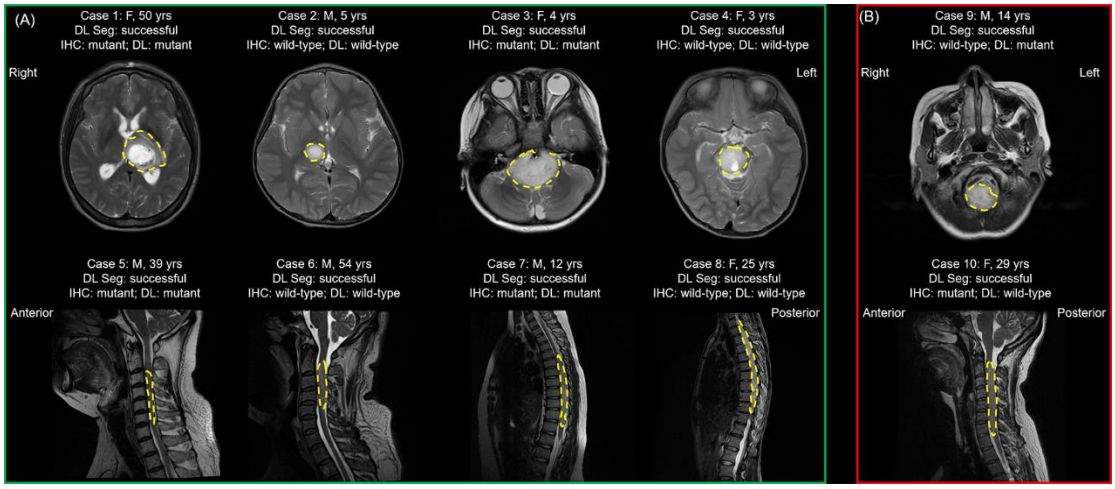


Figure 3. Representative cases of segmentation and H3 K27M status prediction in diffuse midline gliomas using an established deep learning

framework. (**A**) Representative correct prediction cases. Case 1: a female adult patient (aged 50 yrs) with H3 K27M-mutant diffuse midline brain glioma located in the thalamus; Case 2: a male pediatric patient (aged 5 years) with wild-type diffuse midline brain glioma located in the thalamus; Case 3: a female pediatric patient (aged 4 yrs) with H3 K27M-mutant diffuse midline brain glioma located in the brainstem; Case 4: a female pediatric patient (aged 3 years) with wild-type diffuse midline brain glioma located in the brainstem; Case 5: a male adult patient (aged 39 yrs) with H3 K27M-mutant diffuse spinal cord glioma located in the cervical cord; Case 6: a male adult patient (aged 54 yrs) with wild-type diffuse spinal cord glioma located in the cervical cord; Case 7: a male pediatric patient (aged 12 yrs) with H3 K27M-mutant diffuse spinal cord glioma located in the thoracic cord; Case 8: a female adult patient (aged 25 yrs) with diffuse spinal cord glioma located in the thoracic cord. (**B**) Representative incorrect prediction cases. Case 9: a male pediatric patient (aged 14 yrs) with wild-type diffuse midline brain glioma located in the brainstem (DL: mutant); Case 10: a female adult patient (aged 29 yrs) with H3 K27M-mutant diffuse spinal cord glioma located in the cervical cord (DL: wild-type). The yellow contour indicates tumor segmentation using nnU-Net. F, female; M, male; IHC, immunohistochemistry; DL, deep learning; Seg, segmentation; yrs, years.

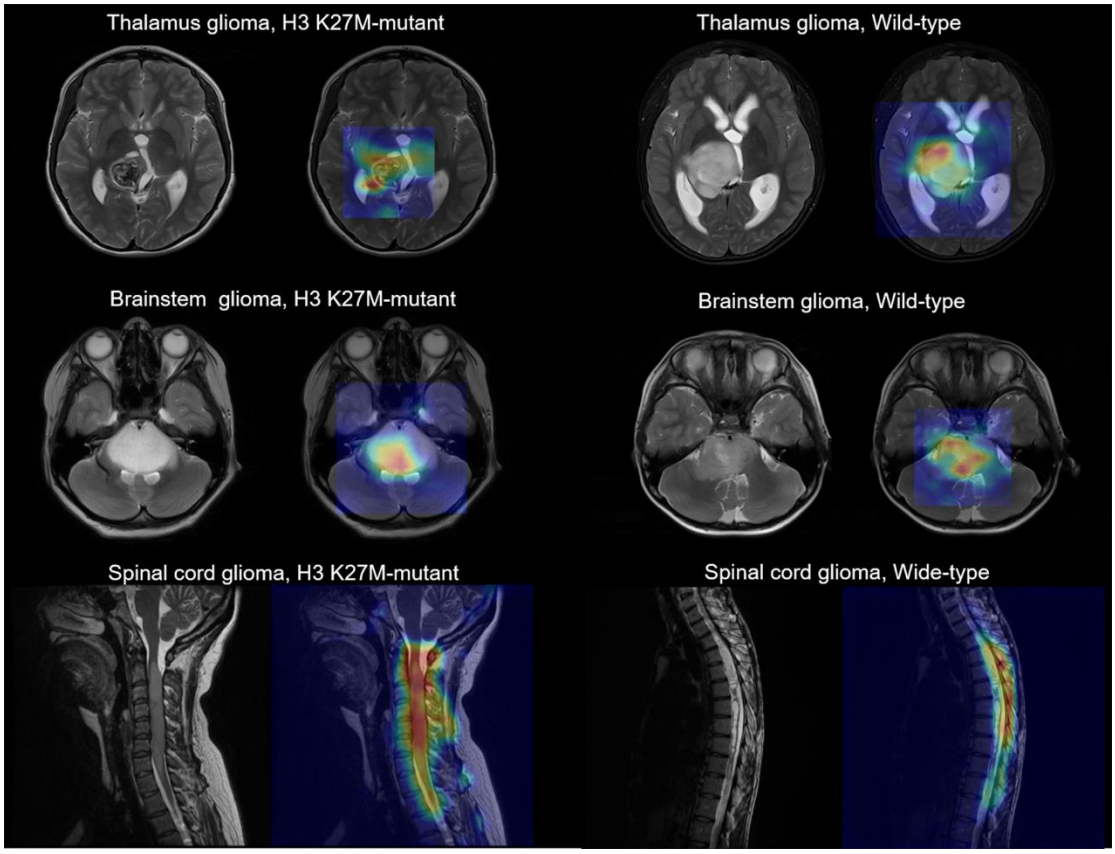


Figure 4. Gradient-weighted Class Activation Mapping (Grad-CAM) of the representative cases for H3 K27M status prediction.

Supplementary materials contents

Supplementary materials

eTables

eTable 1. MR imaging acquisition parameter

Axial T2W images	center-1					center-2					center-3			Sagittal T2W images	center-1				
	GE		Siemens		Philips	GE		Siemens		Philips	GE	Siemens			GE		Siemens		Philips
	HDxt	Discovery MR750	Verio	Prisma	Ingenia CX	Excite	Avanto	TrioTim	Skyra	Achieva	Discovery MR750	TrioTim	Essenza		HDxt	Discovery MR750	Verio	Prisma	Ingenia CX
TR	5160-6000	2687-12248	1800-6000	5020-5550	2800-4600	4000	4100	2500-3000	4500	3000-4000	2809-5752	4000	3000-4000	TR	2140-3060	1000-3268	1800-3000	3000	2000
TE	103-118	89-126	95-99	105-117	87-135	107	93	350-354	105	80-100	108-94	98	87-109	TE	110-122	48-130	94-110	91	110
FA	90	111-142	120-160	90-150	90	90	150	120	150	90	112-142	120	150	FA	90	111-142	120-160	120	90
slice thickness	5-6	3-6	3-5	5	5-6	6	5	1	5	6	4-5	5	5	slice thickness	3	3-5	3	3	3
matrix size	512×512	512×512	256-640×256-640	320-448×320-448	448-576×448-576	512×512	320×288	512×512	448×378	560-640×560-640	512×512	512×416	384-336	matrix size	512×512	512×512	640×640	384×384	512×512

Note: TR, repetition time; TE, echo time; FA; flip angle.

eTable 2. Inter-rater agreement by Cohen's Kappa between two radiologists

	Numbers (%), rater 1	Numbers (%), rater 2	K value
Diffuse midline brain glioma			
Center-1 testing dataset			
Location			
Corpus callosum	78(23%)	78(23%)	1
Thalamus	88(26%)	88(26%)	1
Brainstem	175(51%)	175(51%)	1
MRI presentation			
Edema	101(30%)	110(32%)	0.90
Cystic/necrosis	194(57%)	196(57%)	0.94
Enhancement	220(65%)	236(69%)	0.87
Hydrocephalus	90(26%)	92(27%)	0.97
Center-2 testing dataset			
Location			
Corpus callosum	8(19%)	8(19%)	1
Thalamus	13(31%)	13(31%)	1
Brainstem	21(50%)	21(50%)	1
MRI presentation			
Edema	18(43%)	17(40%)	0.85
Cystic/necrosis	20(48%)	21(50%)	0.86
Enhancement	27(64%)	26(62%)	0.74
Hydrocephalus	9(21%)	9(21%)	1
Center-3 testing dataset			
Location			
Corpus callosum	13(37%)	13(37%)	1
Thalamus	8(23%)	8(23%)	1
Brainstem	14(40%)	14(40%)	1
MRI presentation			
Edema	15(43%)	11(31%)	0.76
Cystic/necrosis	22(63%)	18(51%)	0.77
Enhancement	22(63%)	24(69%)	0.87
Hydrocephalus	7(20%)	7(20%)	1
Diffuse spinal cord glioma			
Center-1 testing dataset			
Location			
Cervical	48(36%)	48(36%)	1
Cervical-thoracic	19(14%)	19(14%)	1
Thoracic	48(36%)	48(36%)	1
Thoracic-lumbar	18(14%)	18(14%)	1
MRI presentation			
Edema	100(75%)	93(70%)	0.87

Cystic/necrosis	60(45%)	51(38%)	0.86
Enhancement	84(63%)	89(67%)	0.92
Cavity	40(30%)	40(30%)	1

Note: n, number.

eTable 3. The prediction of H3 K27M-mutation in diffuse midline gliomas using conventional T2W images by the ensembled predictive outcomes of five DL networks, and performances of conventional machine learning algorithms for H3 K27M-mutation prediction using demographics and conventional MRI features

		Accuracy (%, [95% CI])	Sensitivity (%, [95% CI])	Specificity (%, [95% CI])	AUC (95% CI)
DL using T2W image	**Averaged predictive probability**				
	Diffuse midline brain glioma				
	Center-1 testing dataset	92.1 (86.5-96.6)	96.3 (90.4-100)	85.7 (72.7-96.8)	0.97 (0.93-1.00)
	Center-2 testing dataset	83.3 (71.4-92.9)	96.0 (87.0-100)	64.7 (40.0-87.5)	0.91 (0.78-1.00)
	Center-3 testing dataset	74.3 (60.0-88.6)	81.8 (57.1-100)	70.8 (52.2-88.0)	0.85 (0.78-0.98)
	Diffuse spinal cord glioma				
	Center-1 testing dataset	85.4 (73.2-95.1)	88.9 (72.2-100)	82.6 (65.2-96.0)	0.83 (0.68,0.96)
Lasso regression using demographics and conventional MRI features	**Averaged predictive probability**				
	Diffuse midline brain glioma				
	Center-1 testing dataset	82.0 (78.9-85.9)	83.3 (78.6-88.6)	80.0 (74.1-88.0)	0.86 (0.82-0.91)
	Center-2 testing dataset	76.2 (69.7-81.8)	92.0 (88.9-100)	52.9 (38.5-66.7)	0.72 (0.63-0.82)
	Center-3 testing dataset	62.9 (53.6-71.4)	63.6 (50.0-77.8)	62.5 (52.6-72.2)	0.65 (0.56-0.72)
	Diffuse spinal cord glioma				
	Center-1 testing dataset	70.7 (62.5-78.1)	83.3 (76.9-92.9)	60.9 (50.0-70.6)	0.83 (0.77-0.91)
Support vector machine (linear kernel) using demographics and conventional MRI features	**Majority voting (3/5)**				
	Diffuse midline brain glioma				
	Center-1 testing dataset	82.0 (78.9-85.9)	83.3 (78.6-88.4)	80.0 (74.1-88.0)	NA

	Center-2 testing dataset	73.8 (66.7-81.8)	84.0 (77.8-94.1)	58.8 (46.2-72.7)	NA
	Center-3 testing dataset	62.9 (53.6-71.4)	54.6 (37.5-71.4)	66.7 (57.9-77.8)	NA
	Diffuse spinal cord glioma				
	Center-1 testing dataset	78.1 (71.9-84.4)	83.3 (76.9-92.9)	73.9 (64.7-83.3)	NA
	Averaged predictive probability				
	Diffuse midline brain glioma				
	Center-1 testing dataset	85.4 (81.7-88.7)	92.6 (90.0-95.6)	74.3 (67.9-82.8)	0.86 (0.81-0.91)
	Center-2 testing dataset	69.1 (60.6-75.8)	92.0 (88.9-100)	35.3 (23.1-46.2)	0.73 (0.63-0.82)
	Center-3 testing dataset	51.4 (42.9-60.1)	72.7 (57.1-88.9)	41.7 (31.6-52.6)	0.64 (0.55-0.74)
	Diffuse spinal cord glioma				
	Center-1 testing dataset	70.7 (62.5-78.1)	77.8 (69.2-87.5)	65.2 (55.6-75.0)	0.81 (0.74-0.88)
Support vector machine (Gaussian kernel) using demographics and conventional MRI features	**Majority voting (3/5)**				
	Diffuse midline brain glioma				
	Center-1 testing dataset	79.8 (76.1-84.5)	79.6 (74.4-85.0)	80.0 (74.1-88.0)	NA
	Center-2 testing dataset	66.7 (60.6-72.7)	80.0 (72.2-89.5)	47.1 (33.3-60.0)	NA
	Center-3 testing dataset	57.1 (50.0-64.3)	54.6 (37.5-71.4)	58.3 (47.4-68.4)	NA
	Diffuse spinal cord glioma				
	Center-1 testing dataset	65.9 (59.4-71.9)	88.9 (83.3-100)	47.8 (36.8-58.8)	NA
	Averaged predictive probability				
	Diffuse midline brain glioma				
	Center-1 testing dataset	84.3 (80.3-88.7)	74.3 (67.9-82.8)	90.7 (87.8-95.2)	0.86 (0.81-0.91)

	Center-2 testing dataset	69.1 (60.6-75.7)	92.0 (88.9-100)	35.3 (23.1-46.2)	0.75 (0.67-0.84)
	Center-3 testing dataset	51.4 (42.9-60.1)	72.7 (57.1-88.9)	41.7 (31.6-52.6)	0.66 (0.56-0.76)
	Diffuse spinal cord glioma				
	Center-1 testing dataset	80.5 (75.0-87.5)	83.3 (76.9-92.9)	78.3 (70.6-88.2)	0.82 (0.76-0.90)
Multilayer perceptron using demographics and conventional MRI features	**Majority voting (3/5)**				
	Diffuse midline brain glioma				
	Center-1 testing dataset	77.5 (73.2-81.7)	75.9 (70.5-81.8)	80.0 (74.1-88.0)	NA
	Center-2 testing dataset	69.1 (60.6-75.8)	72.0 (63.2-83.3)	64.7 (53.9-76.9)	NA
	Center-3 testing dataset	51.4 (42.9-60.7)	36.4 (22.2-50.0)	58.3 (50.0-68.4)	NA
	Diffuse spinal cord glioma				
	Center-1 testing dataset	65.9 (59.4-71.9)	83.3 (76.9-92.9)	52.2 (41.2-63.2)	NA
	Averaged predictive probability				
	Diffuse midline brain glioma				
	Center-1 testing dataset	83.2 (78.9-87.3)	74.3 (67.9-82.8)	88.9 (85.4-93.0)	0.87 (0.83-0.91)
	Center-2 testing dataset	69.1 (60.6-75.8)	92.0 (88.9-100)	35.3 (23.1-46.2)	0.74 (0.65-0.82)
	Center-3 testing dataset	57.1 (50.0-64.3)	72.7 (57.1-88.9)	50.0 (40.0-61.1)	0.61 (0.51-0.70)
	Diffuse spinal cord glioma				
	Center-1 testing dataset	65.9 (59.4-71.9)	88.9 (83.3-100)	47.8 (35.3-58.8)	0.82 (0.76-0.91)

Note: AUC, area under the curve; CI, confidence interval; DL, deep learning; NA, not available.

eTable 4. Results from survey-based assessment of Grad-CAM

Assessment target	Survey question	Numbers (%), rater 1	Numbers (%), rater 2	K value
Diffuse midline brain glioma	***Which one is the main activation area in the brain***			
	1.tumor	151 (90.96)	145 (87.35)	0.29
	2.Nontumor brain tissue	15 (9.04)	21 (12.65)	
	Which one is the main activation area within the tumor			
	1.Tumor margin/periphery	78 (46.99)	73 (43.98)	0.85
	2.Central area of the tumor	55 (33.13)	57 (34.34)	
	3.Entire tumor area	33 (19.88)	36 (21.69)	
	Which one is the main activation area outside of tumors			
	1.The activation area does not include the nontumor area at all	0 (0)	0 (0)	0.40
	2.Peritumor area along the tumor margin	113 (68.07)	120 (72.29)	
	3.Brain tissue separated from the tumor	53 (31.93)	46 (27.71)	
Diffuse spinal cord glioma	***Which one is the main activation area in the spinal cord***			
	1.tumor	41 (100)	41 (100)	1
	2.Nontumor spinal cord tissue	0	0	

Which one is the main activation area within the tumor			
1.Tumor margin/periphery	2 (4.88)	3 (7.32)	0.90
2.Central are of the tumor	3 (7.32)	3 (7.32)	
3.Entire tumor area	36 (87.80)	35 (85.37)	
Which one is the main activation area outside of tumors			
1.The activation area does not include the nontumor area at all	0	0	1
2.Peritumor area along the tumor margin	41 (100)	41 (100)	
3.Spinal cord tissue separated from the tumor	0	0	

Note: Kappa statistics were calculated in terms of the interrater agreement between rater 1 (L.Q) and 2 (T.S).

eTable 5. The subgroup analyses of predicting H3 K27M-mutation in diffuse midline gliomas by majority voting

	Dataset (number of mutant and wildtype gliomas)	Accuracy (%, [95% CI])	Sensitivity (%, [95% CI])	Specificity (%, [95% CI])
Pediatric (aged <18 yrs)	***Diffuse midline brain glioma***			
	Center-1 testing dataset (24 vs 9)	87.9 (75.8-97.0)	100 (100-100)	55.6 (20.0-88.9)
	Center-2 testing dataset (7 vs 2)	100 (100-100)	100 (100-100)	100 (100-100)
	Center-3 testing dataset (3 vs 5)	100 (100-100)	100 (100-100)	100 (100-100)
	Diffuse spinal cord glioma			
	Center-1 testing dataset (6 vs 8)	78.6 (57.1-100)	83.3 (66.7-100)	75.0 (37.5-100)
Adult (aged ≥18 yrs)	***Diffuse midline brain glioma***			
	Center-1 testing dataset (30 vs 26)	94.6 (87.5-100)	96.7 (88.9-100)	92.3 (80.8-100)
	Center-2 testing dataset (18 vs 15)	87.9 (75.8-97.0)	94.4 (81.8-100)	80.0 (57.1-100)
	Center-3 testing dataset (8 vs 19)	81.5 (66.7-96.3)	87.5 (57.1-100)	79.0 (58.8-95.2)
	Diffuse spinal cord glioma			
	Center-1 testing dataset (12 vs 15)	88.9 (74.1-100)	91.7 (72.7-100)	86.7 (66.7-100)
DMGs locating in	Center-1 testing dataset	90.0 (76.7-100)	91.7 (72.7-100)	88.9 (72.2-100)

corpus callosum	(12 vs 18)			
	Center-2 testing dataset	100 (100-100)	100 (100-100)	100 (100-100)
	(2 vs 6)			
	Center-3 testing dataset	92.3 (76.9-100)	100 (100-100)	90.0 (70.0-100)
	(3 vs 10)			
DMGs locating in	Center-1 testing dataset	93.2 (86.4-98.3)	97.6 (92.3-100)	82.4 (61.5-100)
thalamus and	(42 vs 17)			
brainstem				
	Center-2 testing dataset	88.2 (76.5-97.1)	95.7 (85.7-100)	72.7 (44.4-100)
	(23 vs 11)			
	Center-3 testing dataset	81.8 (63.6-95.5)	87.5 (60.0-100)	78.6 (55.6-100)
	(8 vs 14)			

Note: yrs, years; CI, confidence interval. CI calculation used a bootstrap method in testing datasets.

eTable 6. Predicting H3 K27M-mutation in diffuse midline gliomas using combination of available T2W-FLAIR and contrast enhanced T1W (cT1W) images by majority voting

	Dataset (number of mutant and wild-type gliomas)	Accuracy (%, [95% CI])	Sensitivity (%, [95% CI])	Specificity (%, [95% CI])
Diffuse midline brain glioma (cases having co-existing T2W, T2-FLAIR and cT1W)				
T2W (n=215*)	Center-1 testing dataset (50 vs 32)	84.2 (75.6-91.5)	86.0 (75.5-95.2)	81.3 (66.7-93.9)
	Center-2 testing dataset (22 vs 17)	76.9 (64.1-89.7)	68.2 (47.8-87.0)	88.2 (70.6-100)
	Center-3 testing dataset (9 vs 20)	79.3 (65.5-93.1)	88.9 (62.5-100)	75.0 (55.0-93.8)
T2W-FLAIR (n=215*)	Center-1 testing dataset (50 vs 32)	82.9 (74.4-90.2)	94.0 (86.5-100)	65.6 (48.7-82.1)
	Center-2 testing dataset (22 vs 17)	69.2 (53.9-82.1)	63.6 (42.3-84.0)	76.5 (54.6-94.4)
	Center-3 testing dataset (9 vs 20)	48.3 (31.0-65.5)	77.8 (44.4-100)	35.0 (15.0-56.5)
cT1W (nr=215*)	Center-1 testing dataset (50 vs 32)	81.7 (73.2-89.0)	98.0 (93.5-100)	56.3 (38.5-73.1)
	Center-2 testing dataset (22 vs 17)	59.0 (43.6-74.4)	72.7 (52.4-90.0)	41.2 (17.7-66.7)
	Center-3 testing dataset (9 vs 20)	41.4 (24.1-58.6)	88.9 (63.6-100)	20.0 (4.7-38.9)
T2W+T2W-FLAIR+cT1W (n=215*)	Center-1 testing dataset (50 vs 32)	81.7 (73.2-90.2)	92.0 (83.7-98.2)	65.6 (48.4-82.4)

	Center-2 testing dataset (22 vs 17)	69.2 (53.9-82.1)	58.8 (33.3-82.4)	77.3 (57.9-94.4)
	Center-3 testing dataset (9 vs 20)	51.7 (34.5-69.0)	55.6 (20.0-87.5)	50.0 (27.8-71.4)
Diffuse midline brain glioma (model trained using maximal available cases having single modalities)				
T2W-FLAIR (n=219*)	Center-1 testing dataset (50 vs 32)	85.4 (78.1-92.7)	96.0 (89.6-100)	68.8 (51.9-84.6)
	Center-2 testing dataset (24 vs 17)	75.6 (61.0-87.8)	91.7 (79.2-100)	52.9 (28.6-76.5)
	Center-3 testing dataset (9 vs 22)	67.7 (51.6-83.9)	86.4 (70.0-100)	22.2 (0-55.6)
cT1W (n=244*)	Center-1 testing dataset (54 vs 35)	80.9 (71.9-88.8)	90.7 (82.5-98.0)	65.7 (48.8-81.1)
	Center-2 testing dataset (23 vs 17)	77.5 (65.0-90.0)	82.4 (61.1-100)	73.9 (54.6-90.9)
	Center-3 testing dataset (11 vs 20)	54.8 (35.5-71.0)	63.6 (33.3-90.9)	50.0 (27.3-72.2)
Diffuse spinal cord glioma (cases having co-existing T2W and cT1W images)				
T2W (n=91*)	Center-1 testing dataset (17 vs 23)	85.0 (83.3-88.9)	82.4 (78.6-87.5)	87.0 (84.2-90.9)
cT1W (n= 91*)	Center-1 testing dataset (17 vs 23)	75.0 (72.2-80.6)	82.4 (78.6-87.5)	69.6 (65.0-76.2)

T2W+cT1W (n=91*)	Center-1 testing dataset (17 vs 23)	82.5 (80.6-86.1)	88.2 (85.7-93.8)	78.3 (73.7-85.0)

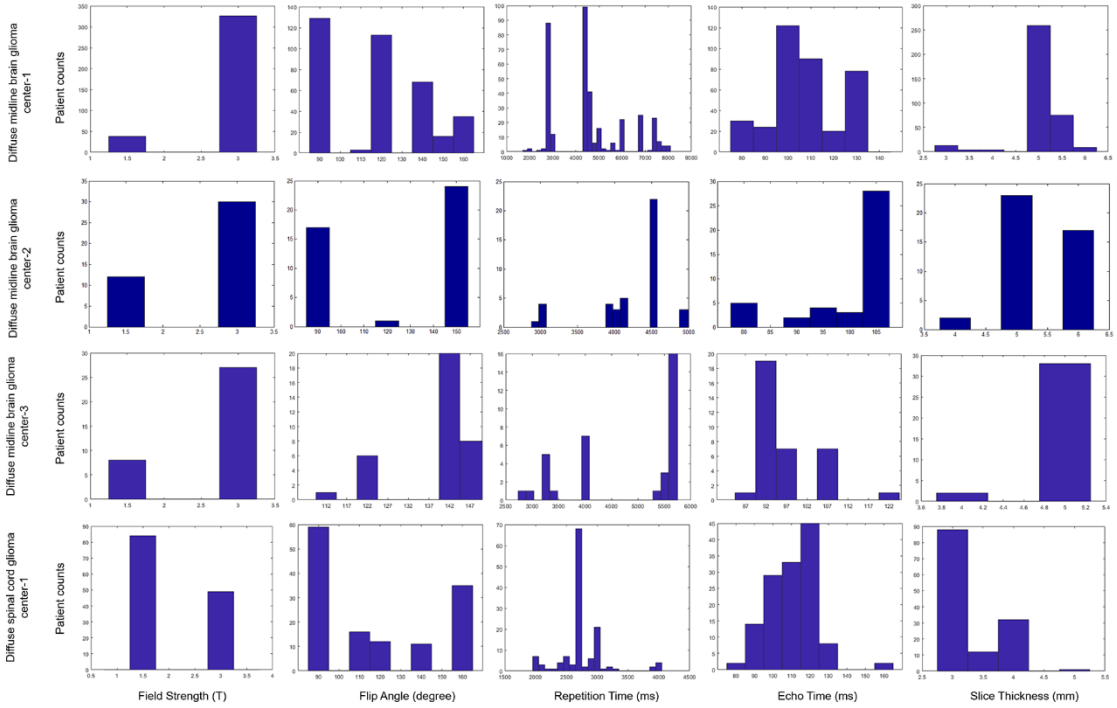
Note: CI, confidence interval. CI calculation used a bootstrap method in testing datasets. Results of diffuse midline brain glioma (model trained using maximal available cases having multiple modalities) was showed for a comparison to determine whether the sample sizes in the train dataset have influence on the H3K27M -mutation prediction using FLAIR and cT1W images. T2W-FLAIR were not available for diffuse spinal cord gliomas in Center-1. *indicates number of cases in training set.

eTable 7. The subgroup analyses of predicting H3 K27M-mutation in diffuse midline gliomas by different magnetic field intensity.

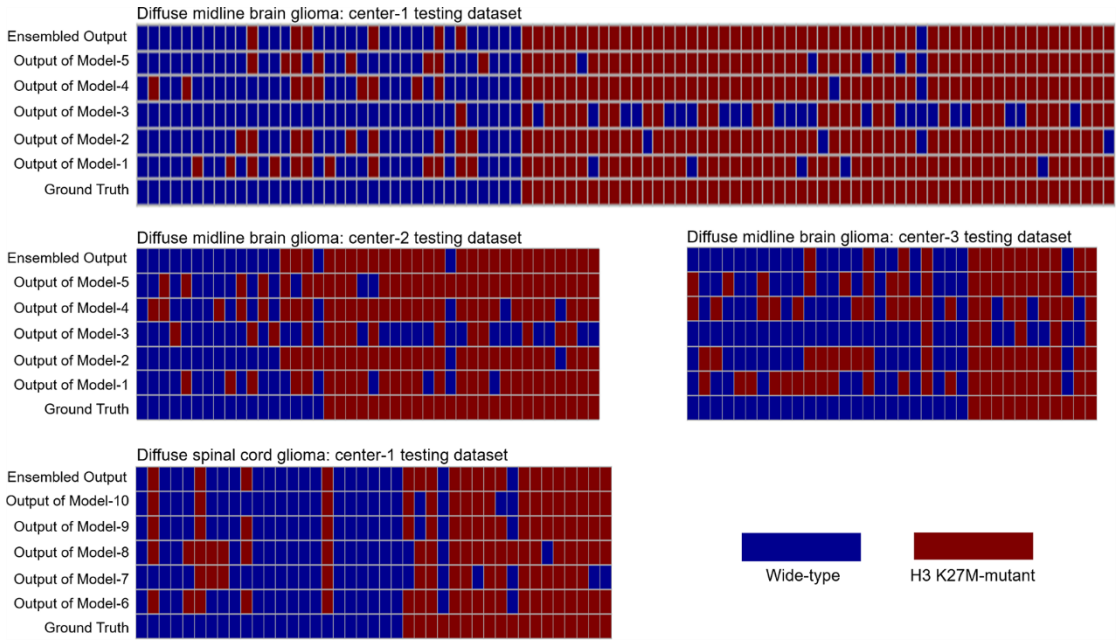
	Dataset (number of mutant and wild-type gliomas)	Accuracy (%, [95% CI])	Sensitivity (%, [95% CI])	Specificity (%, [95% CI])	AUC (95% CI)
1.5T MRI	***Diffuse midline brain glioma***				
	Center-1 testing dataset (6 vs 1)	100 (100-100)	100 (100-100)	100 (100-100)	1
	Center-2 testing dataset (3 vs 2)	80.0 (40.0-100)	100 (100-100)	50.0 (0-100)	0.75 (0.50-1)
	Center-3 testing dataset (0 vs 8)	87.5 (62.5-100)	NA	87.5 (62.5-100)	NA
	Diffuse spinal cord glioma				
	Center-1 testing dataset (13 vs 13)	88.5 (73.1-100)	92.3 (73.3-100)	84.6 (60.0-100)	0.88 (0.74-1)
3.0T MRI	***Diffuse midline brain glioma***				
	Center-1 testing dataset (48 vs 34)	91.5 (85.4-97.6)	95.9 (89.6-100)	84.9 (71.4-96.4)	0.90 (0.83-0.96)
	Center-2 testing dataset (22 vs 15)	91.9 (81.1-100)	95.5 (84.6-100)	86.7 (66.7-100)	0.91 (0.79-1)
	Center-3 testing dataset (11 vs 16)	85.2 (70.4-96.3)	90.9 (71.4-100)	81.3 (60.0-100)	0.86 (0.71-0.97)
	Diffuse spinal cord glioma				
	Center-1 testing dataset (5 vs 10)	80.0 (60.0-100)	80.0 (40.0-100)	80.0 (50.0-100)	0.80 (0.71-0.96)

Note: CI, confidence interval.

eFigure legends



eFigure 1. MRI acquisition details of the T2W images for diffuse midline brain and spinal cord gliomas.



eFigure 2. Predictive model performance by majority voting (⩾3/5) for each diffuse midline glioma patient in testing datasets.

eDocument 1: DL Network for Tumor Segmentation

Network Details

Two separate networks were trained for brain and spinal cord tumor segmentations using 3D nnU-Net. The network architecture of nnU-Net is displayed in **Figure 2**. It follows a 3D U-Net architecture, consisting of an encoder and a decoder, which are interconnected by skip connections. The nnU-Net approach utilizes large patch sizes with small batch sizes with GroupNorm. Input patch sizes of 16×320×320 and 10×416×416 were used for brain and spinal cord tumors respectively. The batch size was set to 2. The inputs were multi-slice axial (brain tumor) or sagittal (spinal cord tumor) T2W images and outputs were multi-slice tumor masks with the same image resolution as the input T2W images.

Training Details

Training objective was the sum of soft Dice and cross-entropy loss, operating on the whole tumor. nnU-Net used stochastic gradient descent with an initial learning rate of 0.01 and a Nesterov momentum of 0.99. Training ran for a total of 200 epochs, where one epoch is defined as the number of training set. The learning rate was set to decay with a polynomial schedule. Training patches were cropped from randomly selected training cases. Data augmentation was applied on the fly during training as those in the original nnU-Net. Segmentation networks were trained using PyTorch, ubuntu 18.04 system with 4 GTX1080TI

GPUs.

Testing Details

Fifty diffuse midline brain gliomas and 19 diffuse spinal cord gliomas were used to quantitatively test the segmentation networks by Dice coefficient.

eDocument 2:DL Model for H3 K27M Mutation Prediction

Network Details

For diffuse midline brain gliomas, the prediction of H3 K27M-mutantion was based on an EfficientNet-B0 network with two convolutional layers, 17 MBConv6 layers, one global pool layer and one fully-connected layer (**Figure 2**). The T2W images were firstly cropped by selecting five slices including the maximum tumor slice and two slices superior to and two slices inferior to the maximal tumor mask slice (according to the segmented tumor mask). Then, the cropped T2W images (five slices) were resampled into a size of 5×224×224 as the input of DL network.

For diffuse spinal cord gliomas, this study adopted a network with a simple architecture including three convolutional layers, two max pool layers, one global pool layer and one fully-connected layer. Given the morphology and MRI acquisition of spinal cord being different from those of brain tumor cases, a different data processing and input size was adopted. The T2W image slices of spinal cord tumors were not cropped but resized into 512×512. Five T2w image slices selected according to the maximal spinal cord tumor slice and its corresponding tumor mask slices (providing tumor location information) were used, resulting in an image of 10×512×512 as the network input. The output was H3 K27M status at the patient-level.

Training Details

Networks were implemented using Pytorch with an adaptive moment estimation optimizer (Adam). The initial learning rate was set to 1e-4 with a batch size of 16 and maximal iterations of 200.

Testing Details

For the diffuse midline brain gliomas, the prediction network was evaluated using an internal prospective independent testing dataset and two external testing datasets by classification accuracy, sensitivity, specificity and area under the curve (AUC).

For the diffuse spinal cord gliomas, the prediction network was evaluated using a prospective independent testing dataset by classification accuracy, sensitivity, specificity and AUC.