scientific reports

OPEN

FA-DeepMSM: a few-shot adapted interpretable multimodal survival model for improved prognostic prediction in glioblastoma

Minyoung Hwang[1], Junhyeok Lee[2], Sihyeon Kim[1], Minchul Kim[3], Seung Hong Choi[4], Sung Soo Ahn[5], Changhee Lee[1,6]✉ & Kyu Sung Choi[4,6]✉

Survival prediction in IDH-wildtype glioblastoma is an inherently time-dependent challenge that remains clinically critical, yet limited by inter-institutional heterogeneity and insufficient labeled data. Existing deep learning prognostic models often overfit to high-dimensional imaging features and fail to generalize across clinical settings. To address these limitations, we developed FA-DeepMSM (Few-shot Adapted Deep Multimodal Survival Model), which integrates self-supervised MRI features with clinical and molecular variables within a few-shot learning framework. High-dimensional MRI embeddings were extracted from 1,359 adult-type diffuse glioma patients using large multi-institutional datasets combining in-house and publicly available cohorts with survival data through a pretrained DINOv2 vision transformer and fused with structured tabular data in a transformer-based survival architecture. The model was fine-tuned under 0-, 5-, 10-, 20-, and 40-shot settings using an external glioblastoma cohort from UPenn ($n = 452$). Few-shot adaptation significantly improved generalization, increasing the average time-dependent C-index from 0.643 (0-shot) to 0.680 (40-shot). To enhance interpretability, FA-DeepMSM incorporates a time-resolved explainability module based on permutation analysis, enabling variable-level risk attribution across survival timepoints (3, 6, 12, 18, and 24 months). Early prognosis was primarily driven by extent of resection, whereas later survival phases were more influenced by MGMT promoter methylation and image-derived features. By addressing data scarcity, cross-modal misalignment, and limited interpretability, FA-DeepMSM establishes a clinically scalable and explainable paradigm for outcome prediction in neuro-oncology.

Glioblastoma is the most aggressive primary brain tumor and is now defined histomolecularly by the 2021 WHO Classification of CNS tumors. Despite the advanced molecular refinements, adult-type diffuse gliomas—especially IDH-wildtype glioblastoma—remain heterogeneous in molecular genetic, and clinical features[1]. This heterogeneity continues to present challenges for prognosis and personalized treatment[2]. Established prognostic factors include clinical variables such as age, performance status, and extent of resection, alongside molecular markers such as O^6-methylguanine-DNA methyltransferase (MGMT) promoter methylation (hereafter, mMGMT) and isocitrate dehydrogenase (IDH) mutation status. In parallel, MRI has emerged as a noninvasive window into tumor biology, potentially encoding phenotypic correlates of underlying molecular heterogeneity[3].

Recent advances in deep learning have enabled the automated extraction of high-dimensional, clinically relevant features from multiparametric MR imaging. Several studies have shown that convolutional and transformer-based models can predict molecular subtypes, tumor segmentation, and tumor grade via multi-task learning, and can even estimate patient survival using imaging data alone[4–8]. However, glioma is inherently a multimodal disease and integrating imaging with clinical and molecular features has become increasingly important for comprehensive prognostic modeling[9–12]. However, although transformer architectures show

[1]Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea. [2]Interdisciplinary Programs in Cancer Biology Major, Seoul National University Graduate School, Seoul, Republic of Korea. [3]Department of Radiology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, South Korea. [4]Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul 03080, Republic of Korea. [5]Department of Radiology, Research Institute of Radiological Science, Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Korea. [6]Changhee Lee and Kyu Sung Choi have contributed equally to this work. ✉email: changheelee@korea.ac.kr; ent1127@snu.ac.kr

promise in integrating heterogeneous modalities, current multimodal models[13,14] often underperform in external validation due to latent modality entanglement, overfitting to high-dimensional image-derived features relative to model capacity, and poor alignment with real-world data distributions[9].

Recent advancements in foundation models have shown promise in tasks such as medical image segmentation by enabling self-supervised learning from large-scale unlabeled data. Adapting self-supervised models trained on such datasets in a zero- or few-shot manner has demonstrated performance comparable to fully supervised approaches, even under limited labeled data conditions, a common challenge in medical domains[15]. However, their application to survival prediction remains largely unexplored. While fine-tuning self-supervised models on task-specific medical applications has become common, adapting them to survival prediction tasks under few-shot conditions remains challenging, especially with large domain gaps across data types. For example, MedSAM demonstrated that self-supervised, few-shot strategies can achieve competitive segmentation performance without large-scale annotated datasets[16–18]. FastGlioma introduced a visual foundation model that detects glioma infiltration from intraoperative optical imaging through few-shot adaptation, enabling rapid, label-free diagnosis[19]. CancerGPT showed that large language models can be adapted with minimal data to accurately predict drug synergy in rare cancer tissues[20]. Despite differences in modality and task, these studies collectively illustrate that self-supervised, few-shot learning enables generalizable and scalable model performance across biomedical domains.

Despite this progress, current multimodal survival prediction approaches face three fundamental limitations. First, the predominant approach involves feeding extracted multimodal embeddings into statistical models such as Cox proportional hazards (CoxPH), which assume proportional hazards and cannot flexibly capture nonlinear, time-dependent survival patterns[21,22]. Second, recent deep learning methods employing complex fusion architectures (e.g., cross-attention mechanisms, transformer-based predictors)[9,23] substantially increase model complexity, elevating overfitting risk in data-scarce survival settings and contributing to poor external validation performance. Third, the scarcity of labeled survival data—particularly across different institutions and rare cancer subtypes—poses fundamental challenges for training robust image encoders and achieving stable model adaptation to new clinical environments.

In this study, we present a multimodal deep learning-based individualized survival model within a few-shot learning framework. The model, termed DeepMSM, is first trained to estimate the risk of cancer progression from adult-type diffuse glioma patients and subsequently is then fine-tuned for survival prediction in glioblastoma (hereinafter referred as FM-DeepMSM). More specifically, the proposed model is first trained on the internal set with relatively larger number of samples and then fine-tuned on the external set under incremental supervision levels (0-, 5-, 10-, 20- and 40-shot) to reflect real-world scenarios where external labels are scarce. The MRI features are extracted in a self-supervised fashion via a vision transformer (ViT) model under knowledge Distillation with NO labels (DINO) v2 [24] from 1,359 adult-type diffuse glioma patients. These variables are then fused with clinical and molecular genetic variables via a modality fusion layer to provide accurate hazard estimation. To further understand the behavior of the proposed model, we incorporate a time-resolved interpretability that quantifies the prognostic contribution of each variable across the survival time, which enables variable-level analysis across heterogeneous glioma populations. By jointly addressing modality gap, generalization failure, and limited interpretability, the proposed framework provides a scalable and time-resolved explainable multimodal survival prediction in neuro-oncology.

Results

Patient characteristics

This study analyzed data from 2,183 patients with histologically confirmed diffuse gliomas across four institutions, grouped into four distinct cohorts. The internal set included 802 patients from Seoul National University Hospital (SNUH; mean age: 55.37 ± 14.95 years; mean OS: 30.67 ± 30.50 months), along with 87 additional patients from the severance (mean age: 58.21 ± 14.11 years; mean OS: 26.48 ± 17.32 months) and 470 patients from the USCF dataset (mean age: 56.76 ± 15.24 years; mean OS: 18.57 ± 17.27 months). External set consisted of 452 patients from UPenn (mean age: 63.69 ± 11.98 years; mean OS: 13.58 ± 11.85 months). Detailed characteristics of these four cohorts are summarized in Table S1.

Performance evaluation: internal validation

In the internal validation, unimodal models using either clinical variables or image alone demonstrated limited performance. The clinical-only deep learning-based survival model (DSM) achieved an average concordance index (C-index) of 0.749 (95% confidence intervals (CI) : 0.742–0.755), while the image-only DSM, based on self-supervised imaging embeddings, yielded a lower performance of 0.697 (95% CI: 0.692–0.702). Traditional models such as CoxPH and Random Survival Forest (RSF), trained on clinical variables alone, achieved C-indices of 0.751 (95% CI: 0.745–0.758) and 0.762 (95% CI: 0.757–0.767), When trained solely on self-supervised imaging embeddings, the performance of CoxPH and RSF declined, with C-indices of 0.692 (95% CI: 0.687–0.697) and 0.723 (95% CI: 0.718–0.728), respectively.

By integrating both modalities, DeepMSM consistently outperformed traditional survival models across most evaluation time points. It achieved the highest average C-index of 0.788 (95% CI: 0.782–0.793), exceeding that of CoxPH (0.760, 95% CI: 0.753–0.767) and RSF (0.771, 95% CI: 0.767–0.775). Statistically significant improvements were observed for DeepMSM at most monthly time points (6–24 months), as well as in the overall average. DeepMSM also achieved the lowest average Brier score of 0.119 (95% CI: 0.118–0.120), further supporting its superior calibration performance (Table 1).

	Time-dependent C-index					
		Eval times				
Model	Data Type	3	6	12	24	48
CoxPH	Clinical	0.799 (0.785, 0.813)	0.722 (0.712, 0.732)	0.737 (0.731, 0.743)	0.749 (0.744, 0.754)	0.749 (0.744, 0.753)
	Image	0.858†† (0.851, 0.866)	0.652 (0.641, 0.663)	0.665 (0.658, 0.672)	0.646 (0.641, 0.651)	0.637 (0.633, 0.648)
	Multimodal	0.849† (0.831, 0.866)	0.732 (0.721, 0.743)	0.733 (0.726, 0.739)	0.749 (0.744, 0.754)	0.738 (0.733, 0.742)
RSF	Clinical	0.820 (0.810, 0.829)	0.756 (0.748, 0.764)	0.744 (0.738, 0.749)	0.757 (0.753, 0.762)	0.734 (0.730, 0.738)
	Image	0.858†† (0.847, 0.868)	0.731 (0.722, 0.740)	0.694 (0.687, 0.701)	0.670 (0.666, 0.675)	0.663 (0.659, 0.667)
	Multimodal	**0.929 (0.923, 0.935)**	0.756 (0.748, 0.765)	0.726 (0.720, 0.732)	0.727 (0.722, 0.732)	0.716 (0.712, 0.720)
Deep MSM	Clinical	0.798 (0.784, 0.812)	0.717 (0.707, 0.728)	0.737 (0.731, 0.743)	0.7511 (0.746, 0.756)	0.740 (0.736, 0.744)
	Image	0.816 (0.803, 0.829)	0.675 (0.666, 0.683)	0.681 (0.675, 0.688)	0.662 (0.657, 0.666)	0.652 (0.648, 0.656)
	Multimodal	0.866 (0.856, 0.877)	**0.778 (0.769, 0.786)**	**0.767 (0.760, 0.773)**	**0.770 (0.765, 0.774)**	**0.757 (0.754, 0.761)**

Table 1. Internal Test Set Performance: Clinical-Only vs. Image-Only vs. Multimodal Models. *P-values from paired t-tests against the Multimodal DeepMSM: No symbol: P-value 0.001; †: 0.001 ≤ P-value 0.01; ††: 0.01 ≤ P-value. Parentheses indicate 95% confidence interval; Bold text represents the highest score; Underlined text represents the second-highest score.*

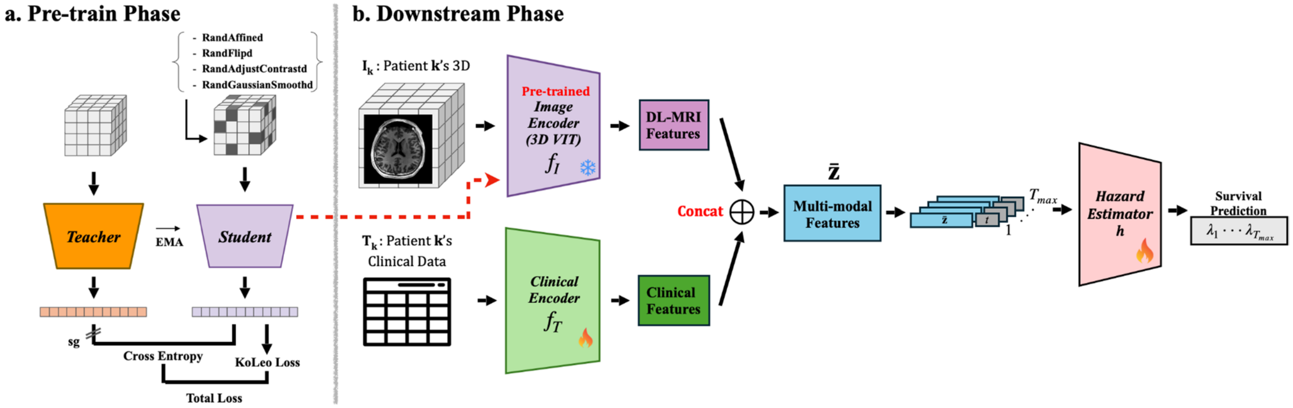


Fig. 1. The Overall Framework: **a** Pretraining using DINOv2 trained on large-scale data to extract self-supervised MRI image features; **b** the proposed multimodal transformer model for survival prediction.

Performance evaluation: GBM-only external validation

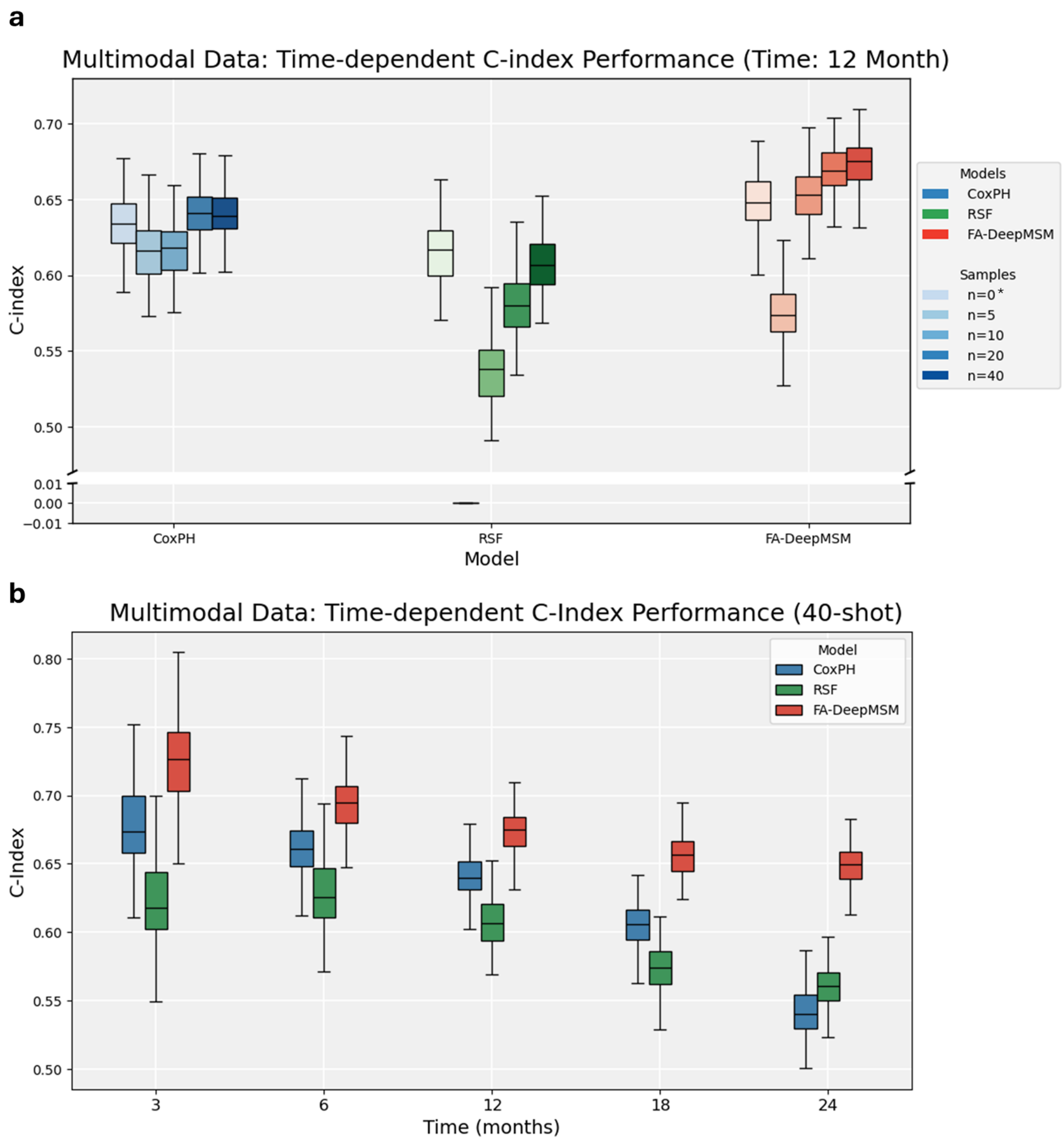
We evaluated the internally-trained models in a zero-shot setting on an external cohort composed exclusively of glioblastoma (GBM) cases. Surprisingly, the performance of the multimodal models—combining imaging and clinical data—was inferior to that of unimodal models using only clinical variables. We hypothesize that this performance degradation stems from the multimodal models being overfitted to the heterogeneous tumor subtypes and biased toward image features present in the internal cohort. In particular, imaging features are more susceptible to domain shifts across institutions and disease subgroups, which may have impaired the generalizability of the learned representations to the GBM-only external cohort (Fig. 1).

Few-shot fine-tuning was performed to adapt the internally-trained model to external GBM data, utilizing 5, 10, 20, and 40 shots. In Cox models, imaging features showed negative or negligible contribution at 5 shots but progressively improved with more data, showing neutral effect at 40 shots. Clinical variables such as Karnofsky Performance Status (KPS), WHO grade, IDH status, and tumor location showed limited but stable contributions. In contrast, deep learning models demonstrated consistent performance gains with increasing shot numbers (Fig. 2a), with imaging features contributing positively from 10 shots onward, highlighting the effectiveness of few-shot adaptation for image–clinical alignment (Table 2, Supplementary Fig. 5). Notably, at the 40-shot setting, the proposed FA-DeepMSM model consistently outperformed CoxPH and RSF across all time intervals (3 to 24 months), as shown in Fig. 2b.

Time-dependent interpretability

Permutation-based feature importance analysis revealed how the prognostic contribution of each clinical variable changes over time, as measured by the time-dependent C-index at each time point (Figs. 3a-b). We observed that extent of resection (EOR) had a stronger prognostic contribution at earlier time points, while MGMT promoter methylation became more important at later stages of follow-up (Fig. 3b).

In our model, MRI-derived deep learning-based prognostic indices (DPIs) initially showed low or negligible importance in the 5-shot setting. However, as the number of adaptation samples increased, the model was



* FA-DeepMSM without fine-tuning (n=0) is equivalent to the original DeepMSM (i.e., zero-shot setting).

Fig. 2. Comparison of Model Discrimination Performance Over Few-shot and Time Settings: **a** C-index performance at the 12-month timepoint across three models (CoxPH, RSF, and FA-DeepMSM), evaluated in both zero-shot ($n = 0$) and varying few-shot fine-tuning scenarios ($n = 5$, 10, 20, 40). **b** Longitudinal comparison of model discrimination performance at 3, 6, 12, 18, and 24 months following 40 shot fine-tuning. The figure illustrates how each model performs at different timepoints, enabling direct comparison of time-specific prognostic discrimination.

able to increasingly leverage image information—reflected by a consistent rise in ΔC-index (i.e., difference in the C-index) for the image feature (Supplementary Fig. 5). In contrast, the Cox model did not exhibit similar recovery of image importance, showing the importance of DPIs maintaining near-zero or even negative ΔC-index values across all shot levels. This highlights a key advantage of deep learning models in capturing transferable representations from imaging data under low data regimes.

	External: UPenn (Time = 12 Month)					
Data Type	**Model**	**0-shot**	**5-shot**	**10-shot**	**20-shot**	**40-shot**
Clinical	**CoxPH**	0.661 (0.657, 0.664)	0.634 (0.631, 0.638)	0.644 (0.640, 0.647)	0.658 (0.655, 0.661)	0.663 (0.660, 0.667)
	RSF	0.668 (0.665, 0.672)	0.436 (0.432, 0.440)	0.571 (0.567, 0.575)	0.618 (0.615, 0.622)	0.642 (0.638, 0.645)
	FA-DeepMSM	0.661 (0.657, 0.665)	0.642 (0.638, 0.645)	0.628 (0.624, 0.631)	0.639 (0.636, 0.643)	0.657 (0.653, 0.661)
Image	**CoxPH**	0.583 (0.579, 0.587)	0.478 (0.474, 0.483)	0.505 (0.501, 0.509)	0.548 (0.543, 0.552)	0.600 (0.596, 0.604)
	RSF	0.605 (0.602, 0.609)	0.441 (0.437, 0.444)	0.536 (0.531, 0.540)	0.563 (0.559, 0.568)	0.591 (0.586, 0.595)
	FA-DeepMSM	0.580 (0.577, 0.584)	0.512 (0.508, 0.516)	0.536 (0.531, 0.540)	0.576 (0.571, 0.580)	0.621 (0.617, 0.625)
Multimodal	**CoxPH**	0.631 (0.627, 0.635)	0.616 (0.611, 0.620)	0.618 (0.614, 0.622)	0.642 (0.638, 0.645)	0.641 (0.637, 0.644)
	RSF	0.611 (0.607, 0.615)	0.456 (0.452, 0.460)	0.539 (0.534, 0.543)	0.582 (0.577, 0.586)	0.607 (0.603, 0.610)
	FA-DeepMSM	0.642* (0.638, 0.646)	0.575 (0.571, 0.579)	0.653 (0.650, 0.658)	0.669 (0.666, 0.673)	0.674 (0.671, 0.678)

Table 2. External Test Set Performance: Across Zero- and Few-Shot Settings (12 months). *P-values from paired t-tests against the Multimodal FA-DeepMSM: Every element has a less than 0.001 P-value. Parenthesis indicates 95% confidence interval; Bold text represents the highest score; Underlined text represents the second-highest score. * FA-DeepMSM without fine-tuning (n = 0) is equivalent to the original DeepMSM (i.e., zero-shot setting).*

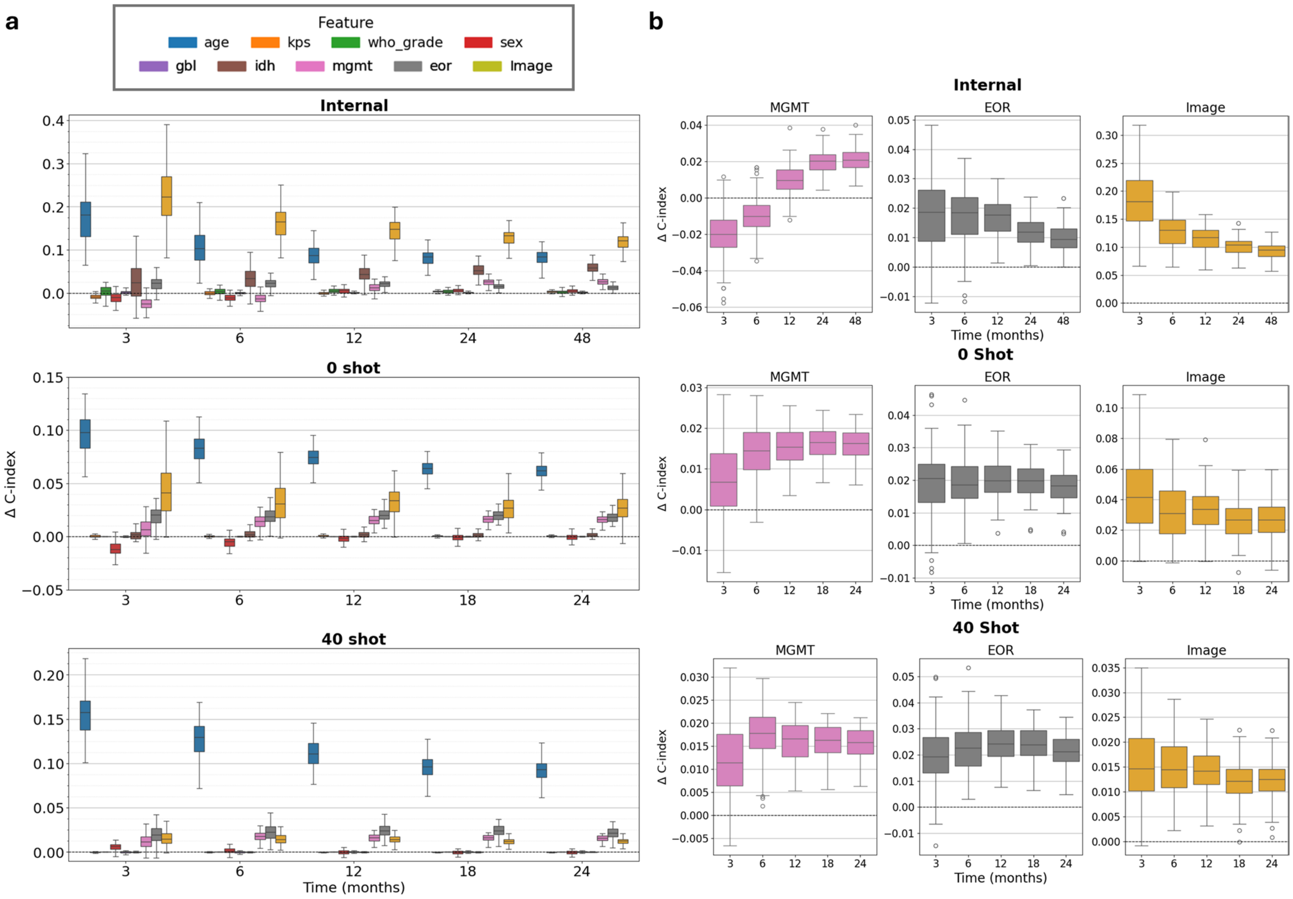
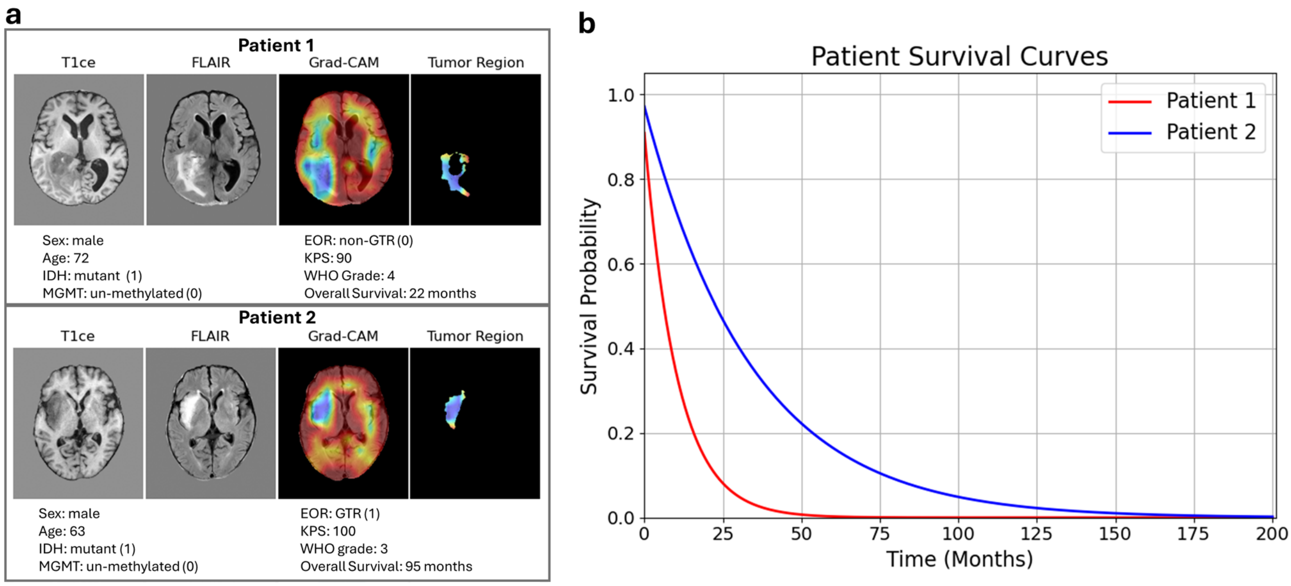


Fig. 3. Longitudinal Shifts in Prognostic Value: **a** Longitudinal analysis of prognostic contributions for all input features, illustrating how their relative importance changes across multiple timepoints in both internal and external cohorts. **b** Focused analysis of three clinically core features illustrates their dynamic prognostic impact across time.

Patient-Level interpretability

To better understand the spatial reasoning of the proposed model, we applied Grad-CAM to visualize the regions that contribute most to the predicted survival risk. In patient-specific interpretation of the comparative representative cases (Fig. 4a), the activation map (third column) highlight regions that closely align with the tumor boundaries (fourth column), as identified by manual segmentation. Interestingly, the highlighted areas predominantly overlap with the enhancing tumor and peritumoral edema without given the tumor segmentation



Patient 1

Sex: male
Age: 72
IDH: mutant (1)
MGMT: un-methylated (0)
EOR: non-GTR (0)
KPS: 90
WHO Grade: 4
Overall Survival: 22 months

Patient 2

Sex: male
Age: 63
IDH: mutant (1)
MGMT: un-methylated (0)
EOR: GTR (1)
KPS: 100
WHO grade: 3
Overall Survival: 95 months

Fig. 4. Patient-Level Interpretation of Deep Multimodal Survival Model via Grad-CAM Visualization: **a** Grad-CAM visualizations for two comparative representative patients, overlayed on T1CE and FLAIR images. The highlighted areas align well with annotated tumor lesions, indicating that the model focuses on clinically relevant tumor lesions when estimating survival risk. **b** Patient-specific prognostic visualization: Predicted survival curves derived from multimodal inputs for two patients. Patient 1 *(red)*, a 63-year-old male with IDH-wildtype glioblastoma in the right temporo-occipital lobe, MGMT-unmethylated, KPS 90, underwent subtotal resection and survived 22 months (predicted probability at death: 10.9%). Patient 2 *(blue)*, a 72-year-old male with IDH-mutant astrocytoma in the right insula, MGMT-unmethylated, KPS 100, also underwent gross total resection and survived 95 months (predicted probability at death: 5.7%). The model predicted a substantially lower survival probability across time for Patient 1, consistent with more aggressive clinical and imaging features.

masks, suggesting that the model leverages clinically relevant features in making its predictions. We further compared individualized survival predictions for the two patients with distinct trajectories of clinical outcomes. Patient 1, a 72-year-old male with glioblastoma (WHO grade 4, KPS: 100, MGMT: un-methylated and EOR: non-total extent), died at 22 months, at which point the model predicted a survival probability of 0.109. Patient 2, a 62-year-old male with slightly more favorable clinical factors (WHO grade 3, KPS: 90, MGMT: un-methylated and EOR: non-total extent), died at 95 months, with a model-estimated survival probability of 0.06 at that time. In both cases, the predicted probabilities at their respective times of death reflect a realistic downward trajectory and align with their relative risk profiles. These differences are clearly visualized in the survival curves in Fig. 4b, where Patient 1's curve declines sharply compared to the more gradual descent observed for Patient 2.

Risk score comparison: multimodal vs. unimodal

In Fig. 5a, the distribution of risk scores on the external cohort reveals notable differences across the deep survival models. The clinical-only model exhibits a relatively flat distribution with scores concentrated in a specific range, indicating limited variability in its predictions. The image-only model shows a narrow distribution, with most risk scores clustered tightly around the median, suggesting low variance and limited risk stratification power. In contrast, the multimodal model demonstrates a broader distribution of risk scores without being overly concentrated in a particular interval, reflecting both high variance and a more continuous risk spectrum. This indicates that the multimodal model (i.e., DeepMSM) not only achieves better discrimination, as also demonstrated in Table 1, and at the same time delivers more reliable calibration in its risk estimates. The ability to distinguish high- and low-risk individuals while maintaining well-calibrated predictions highlights the strength of integrating clinical and imaging features, particularly in few-shot scenarios. This pattern is further supported by Fig. 5b, where the separation between the survival curves of the high-risk and low-risk groups is noticeably more pronounced in the multimodal model compared to the unimodal models.

Method and materials

Patient cohorts

In this retrospective multicenter study, we analyzed data from 1,871 patients with histologically confirmed adult-type diffuse glioma according to 2021 WHO classification of CNS tumors (i.e., IDH-wildtype glioblastoma, IDH-mutant, 1p/19q-nondeleted astrocytoma, IDH-mutant, 1p/19q-codeleted oligodendroglioma), collected across four independent institutions. For model development, we used a cohort from Seoul National University Hospital (SNUH; $n = 802$, collected between May 2012 and December 2021), Severance Hospital ($n = 87$, collected between December 2016 and December 2017) and the University of California, San Francisco (UCSF-PDGM;

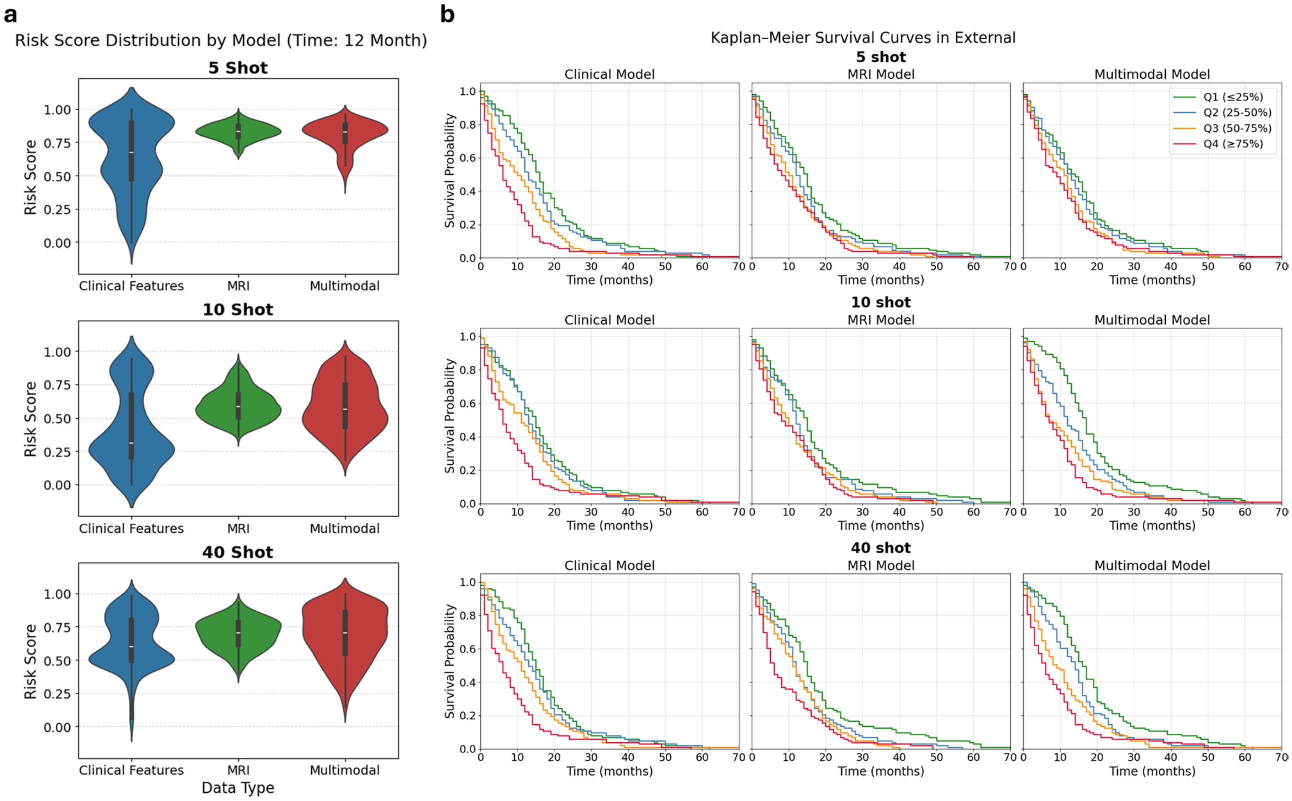


Fig. 5. Evaluation of Model Calibration and Risk Stratification Performance under Few-shot Settings: (a) Risk score distributions across models (12 month, 10/20/40 shot). (b) Model-wise Kaplan–Meier survival curves stratifying patients into low and high-risk groups.

$n = 470$, publicly available). For external validation, we employed an independent cohort from the University of Pennsylvania (UPenn-GBM; $n = 452$, publicly available). Patients were included only if they possessed all essential MRI sequences required for tumor segmentation and analysis and had complete overall survival information. All cohorts received standard-of-care treatment, including initial tumor resection followed by chemotherapy and/or radiotherapy, when indicated. All experimental protocols were approved by the Institutional Review Boards of SNUH and Severance Hospital, and the Institutional Review Boards of SNUH and Severance Hospital waived the need for informed consent. All procedures followed the principles of the Declaration of Helsinki. Analyses of the de-identified public datasets (UCSF-PDGM and UPenn-GBM) were exempt from IRB approval under institutional policies for secondary use of anonymized data. Overall survival time was calculated from the date of initial diagnosis to death attributed to the disease or the last follow-up. We trained the model on diffuse glioma due to greater data availability, but intentionally validated it on IDH-wildtype glioblastoma (GBM)—the primary population of prognostic interest—using the UPenn-GBM cohort as an optimal external test set.

Clinical and genetic variables—including age, sex, KPS, EOR, WHO grade, glioma pathology type, IDH mutation status, and MGMT status—were collected for survival prediction.

While the internal cohorts (i.e., SNUH, Severance, UCSF) exhibited similar distributions in clinical variables and survival time, these distributions were statistically different from those in the UPenn cohort (Table 2, Supplementary Materials Fig. 1). This heterogeneity led us to designate the UPenn dataset as the external validation cohort and the combined data from the other three centers as the internal cohort for model training and validation.

MRI preprocessing and image feature extraction

MRI data were acquired using distinct protocols across participating institutions (detailed acquisition parameters including scanner specifications, field strengths, contrast agents, and sequence parameters are provided in Supplementary 1).

For each patient, four MRI sequences were used: T1-weighted (T1), contrast-enhanced T1 (T1CE), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR). All scans were skull-stripped, resampled to 1 mm isotropic resolution, co-registered to a common anatomical template, and intensity normalized.

As illustrated in Fig. 1, the vision backbone for extracting DPI employed the vision transformer (ViT)[25] architecture, which has demonstrated strong performance across a range of imaging tasks, particularly when capturing global dependencies and contextual information is crucial. Our vision backbone was adapted to handle multiparametric 3D MRI and further trained in a self-supervised manner on the internal set via self-DIstillation with NO labels (DINO) v2. For each 3D MRI, the sequence-wise averaged embeddings were utilized as the DPI, providing semantically rich encoding of tumor characteristics without requiring manual labeling and demonstrating strong transferability across heterogeneous clinical datasets. Additional implementation details are provided in Supplementary 1 and Table S2.

Clinical variables and missing data handling

Structured clinical variables included patient age, KPS, EOR, and IDH mutation status. For tabular clinical variables, missing values were imputed using the mean (for continuous variables) or mode (for categorical variables) calculated from the available samples within the same dataset. In cases where a particular clinical variable was entirely missing across all samples in a dataset (e.g., KPS in the UCSF cohort), we imputed missing values using the mode or mean from another internal cohort with similar clinical characteristics. The integration of clinical and genetic variables into the survival prediction model involved preprocessing numerical variables, where age was standardized using Z-scores and KPS was scaled to a [0, 1] range using min-max normalization. Binary variables such as EOR, IDH, and MGMT, as well as the ordinal variable WHO grade, were used in their original form without one-hot encoding.

Model architecture and survival prediction

The proposed Deep Multimodal Survival Model (DeepMSM) is a multi-modal deep learning-based survival model which learns the distribution of time-to-event outcomes to assess the individualized risk of cancer progression utilizing both 3D MRIs and clincal variables. In this study, the time-to-event outcome indicates the time at which cancer-specific death occurs, or the time at which a patient is right-censored (i.e., lost to follow-up). DeepMSM consists of three key components as illustrated in Fig. 1: the 3D-ViT vision backbone for extracting the DPIs from MRIs (i.e., the image encoder), a modality fusion layer for concatenating the embedding from the image encoder and the clinical variables, and a survival prediction head (i.e., an MLP with a sigmoid function) that integrates embeddings from the MRIs and those from clinical variables to provide time-dependent conditional hazard estimates (i.e., the hazard network). These conditional hazard estimates are then accumulated to calculate the individualized survival predictions. The model is initially trained based on the negative log-likelihood loss function to account for right-censored patients. To train our network to our relatively small dataset with high-dimensional 3D MRIs and tabular-based clinical variables, we regularize the network by freezing the pretrained 3D-ViT vision backbone and by sharing the same hazard network across different time horizons, as suggested in Lee et al.[26].

Few-shot learning for external validation

To simulate data-limited clinical scenarios for external validation, our model trained on the internal training set was externally validated under five supervision levels: 0-shot which indicates the pretrained model is directly applied for external validation, and 5-, 10-, 20-, and 40-shot which indicates finetuning the pretrained DeepMSM with a small number of labeled samples (shots) in the external training set ($n = 40$). We denote the few-shot finetuned version of DeepMSM as **FA-DeepMSM** (i.e., *Few-shot Adapted DeepMSM*). For efficient adaptation, we adopt a linear probing approach, in which only the last projection layer of the survival prediction head is fine-tuned, with all preceding layers kept frozen. This strategy is applied to a model pretrained on a large-scale internal dataset.

More specifically, we performed the following procedure: we first randomly selected 40 patients for few-shot training from the entire external cohort ($n = 452$), using the remaining held-out 412 patients as the external testing set. This few-shot approach was used to help the models trained with the internal data generalize well to out-of-distribution samples—which included only *GBM* cases collected from different hospitals/countries. We evaluated two external validation scenarios: (i) zero-shot scenario, where the internally trained models are directly applied without fine-tuning, and (ii) few-shot scenario, where we randomly selected 5, 10, 20, or 40 external samples to fine-tune the internally-trained model. Notably, the traditional ML models cannot be fine-tuned and thus are retrained from scratch on the few-shot samples, using the identical model configurations and hyperparameters from the internally-trained models.

Machine learning-based survival prediction

For comparison, we utilized the Cox proportional hazards (Cox PH) model[27] and the random survival forest (RSF) model[28], an ensemble of survival trees which estimates the cumulative hazard function without the proportional hazard assumption. These baseline survival models utilize both MRI-derived DPIs from the trained vision backbone, and clinical and genetic variables—including age, sex, KPS, EOR, WHO grade, glioma pathology type, IDH mutation status, and MGMT methylation status —for survival prediction.

Evaluation metrics and ablation strategy

Given our model's ability to capture the time-varying effect of MRIs and clinical variables, we employed the time-dependent C-index[29] and time-dependent Brier scores[30] for assessing discrimination performance and calibration performance, respectively, at multiple clinically meaningful evaluation time horizons. We conducted modality ablation studies based on the proposed deep learning-based survival model (DSM) to quantify the contribution of each input source (image-only, clinical-only, and combined). This involved training distinct deep survival models based on our proposed architecture, with each model exclusively using features from its corresponding modality (i.e., only image or only clinical inputs, respectively).

For model development and internal validation, we split the internal data into a random 80/20 training/testing split, i.e., 80% of the study population ($n = 1087$) for training and the remaining 20% ($n = 272$) held out for testing. For hyper-parameter optimization and early stopping, we further split the training data into 75/25 training/validation split, which gives 815 training samples and 272 validation samples. To ensure robust evaluation, we performed bootstrapping of 272 patients on the testing set with 100 iterations. For this cohort, the time-dependent C-index and Brier score were assessed at 3, 6, 12, 24, and 48 months since diagnosis.

For external validation, the performance evaluation was conducted on the same external testing set ($n=412$) for both zero-shot and few-shot scenarios. We performed bootstrapping of 412 patients on the external testing set with 100 iterations. Consistent with the internal validation, we used the time-dependent C-index and Brier score, but adjusted the evaluation time horizons to 3, 6, 12, 18 and 24 months to reflect the shorter survival times observed in the external cohort. The effectiveness of few-shot adaptation was analyzed by examining performance trends across different shot levels.

Time-dependent interpretability

To interpret model outputs, we implemented a post-hoc time-dependent interpretability module based on the permutation-based variable importance. In particular, for each tabular clinical variable and MRI-derived DPI embeddings, we randomly permuted across individuals and quantified the performance drop in the time-dependent C-index of the trained model resulting from the broken relationship between the permuted feature and the true outcome. Based on these time-dependent interpretability modules, we estimated the dynamic prognostic relevance of each feature over survival time. This allowed us to visualize when and how specific variables—such as IDH mutation or T1CE embedding—contributed most to predicted risk, enabling us to understand the time-varying effect of each feature on survival predictions.

Statistical analysis

Model performance was assessed using the time-dependent C-index and the time-dependent Brier score with 95% CI, respectively. The performance of the multimodal survival models was compared against naive survival models trained solely on imaging or clinical data alone. Statistical significance of the performance improvements was conducted using a paired Student's t-test. To evaluate the prognostic efficacy of the DPI, Kaplan-Meier survival analysis.

All statistical analyses were conducted using Python with the *scipy*, *torchsurv*, *sklearn*, *sksurv*, and *lifelines* libraries. A p-value of less than 0.05 was considered statistically significant, and all CIs were reported at the 95% level.

Discussion

We developed a few-shot multimodal survival prediction framework for diffuse glioma using self-supervised image embeddings and clinical variables. By extracting high-dimensional MRI features with DINOv2-pretrained vision transformers (Fig. 1a), we integrated them with clinical data for survival prediction under low-data regime (Fig. 1b). Our findings show that few-shot adaptation improves image feature contribution to survival prediction, highlighting the potential of foundation vision models in neuro-oncology. Time-dependent interpretability analysis revealed variable-level prognostic importance with the model generalizing well from diffuse glioma to glioblastoma, improving performance despite distribution differences.

While recent multimodal models show promise, their performance often degrades in external validation due to data distribution shifts and overfitting to high-dimensional imaging features. In our analysis, the clinical-only model exhibited a 11.1% relative drop in C-index (from 0.749 to 0.666), whereas the multimodal model suffered an even greater 18.4% decrease (from 0.788 to 0.643) when applied to the external cohort. Models using only clinical variables or conventional machine learning often outperform multimodal models on external cohorts[31]. This performance gap is largely attributable to the particular vulnerability of imaging features to domain shifts such as scanner variability and acquisition protocol differences[32,33]. To mitigate these limitations and fully leverage the representational power of multimodal models, we propose a few-shot adaptation strategy that aligns imaging and clinical modalities in a data-efficient manner.

Our strategy fundamentally diverges from existing multimodal approaches to resolve these specific limitations. The predominant paradigm in current multimodal survival analysis involves feeding extracted deep embeddings directly into statistical models such as CoxPH[21,22]. However, this approach forces high-dimensional, complex multimodal features into a framework constrained by the proportional hazards assumption, limiting the ability to model non-linear and time-varying dynamics. Conversely, recent complex fusion architectures (e.g., cross-attention[9], transformer predictors[23] swing to the opposite extreme, dramatically increasing complexity and amplifying overfitting risks in data-scarce settings. In contrast, our approach prioritizes parameter efficiency. We achieve this by coupling a lightweight neural hazard network—which inherently handles non-linearity unlike CoxPH—with a frozen, self-supervised pretrained encoder. This architectural decision minimizes trainable parameters, directly mitigating the risk of overfitting while facilitating robust adaptation to new clinical environments.

A key contribution of our study is demonstrating that self-supervised MRI embeddings can generalize across multiple institutions and scanner protocols, even with a small number ($n=40$) of labeled training cases. Previous glioma prognostication studies, including multimodal approaches, have relied on fully supervised methods or handcrafted radiomic features, requiring extensive annotation and often failing to generalize to out-of-distribution samples. In contrast, our approach captures latent imaging phenotypes efficiently, showing consistent predictive performance improvements across internal and external datasets. Notably, the model maintained robust performance in IDH wild-type glioblastoma, a clinically challenging subgroup with molecular heterogeneity. In the internal set, the clinical-only deep learning model achieved a time-dependent C-index of 0.749 (95% CI: 0.742–0.755), while the image-only model with self-supervised embeddings scored 0.697 (95% CI: 0.692–0.702), indicating that imaging features alone capture meaningful but incomplete prognostic information. In contrast, DeepMSM, integrating both clinical and MRI data, achieved the highest C-index of 0.788 (95% CI: 0.782–0.793), emphasizing the complementary nature of tabular and imaging modalities. Multimodal CoxPH and RSF models yielded C-indices of 0.760 (95% CI: 0.753–0.767) and 0.771 (95% CI: 0.767–0.775), respectively (Table 1). These results demonstrate that the proposed deep learning–based model is able to effectively capture long-term survival patterns by integrating imaging and tabular clinical variables. In the external GBM cohort, the DeepMSM model without adaptation (0-shot) achieved an average C-index of 0.643 (95% CI: 0.638–0.647), highlighting performance limitations under domain shift. In contrast, FA-DeepMSM, trained with only 40 labeled samples from the target domain, significantly improved the C-index to 0.680 (95% CI: 0.676–0.683), a 5.7% relative improvement, clearly demonstrating the effectiveness of minimal adaptation for cross-institutional generalization (Table 2; Fig. 2b). FA-DeepMSM began to surpass its zero-shot baseline from the 10-shot setting onward, with performance consitantly increasing as more adaptation samples were used—unlike CoxPH, whose performance remained largely unchanged, or RSF, which consistently exhibited lower performance (Fig. 2a). In addition, FA-DeepMSM consistently outperformed the unimodal models using only image or those using only clinical variables, demonstrating the strength of integrating multimodal information in achieving higher discriminative power even when applied to the external cohort. Specficially, in the 40-shot setting, the tabular-only model achieved a C-index of 0.657 (95% CI: 0.653–0.661), while the image-only model showed a lower performance of 0.621 (95% CI: 0.617–0.625). In contrast, the multimodal FA-DeepMSM achieved the highest performance with a C-index of 0.674 (95% CI: 0.671–0.678) (Table 2).

Despite growing interest in multimodal deep learning, recent studies highlight its vulnerability to domain shifts. For instance, Silva-Rodríguez et al. (2024)[34] found that multimodal models frequently failed to generalize across institutions, even when using sophisticated pretraining. To address this, we deployed our few-shot adaptation framework based on a linear probing strategy that fine-tunes only the final layer of the hazard estimator (Fig. 1b). This mitigates overfitting due to the relatively large model capacity in low-data regime[35]. Our ablation study on few-shot adaptation shows that the proposed framework can overcome early-stage inter-modal misalignment (i.e., between image and clinical variables) when applied to the external cohort. Specifically, at low shot levels ($n = 0$, 5, and 10), clinical variables alone dominated survival prediction, while imaging features either contributed negatively to performance or remained neutral. However, as the number of labeled samples increased (n = 20, 40), the model learned to exploit imaging representations more effectively, with consistent C-index gains attributable to image features from 10 to 40 shots. This suggests that our few-shot adaptation effectively aligns heterogeneous modalities in a clinically coherent manner even with a small number of – samples[34,35].

A key clinical implication of this work is its ability to accommodate the considerable practice variability across neuro-oncology centers. Differences in the extent of resection (e.g., supramaximal versus conventional resection), access to anti-VEGF agents such as bevacizumab or cediranib, adoption of immune checkpoint inhibitors (e.g., PD-1/PD-L1 blockade), and the selective use of targeted agents such as vorasidenib for non-enhancing IDH-mutant gliomas collectively create center-specific survival patterns that hinder the generalizability of externally trained models. By leveraging self-supervised image embeddings extracted from large heterogeneous public benchmarks via a 3D ViT–based DINOv2 backbone, the proposed framework maximizes the prognostic value contained in MRI while maintaining robustness under substantial domain variability. By enabling local calibration with only a small number of institution-specific samples, the few-shot adaptation strategy in FA-DeepMSM provides a practical means of constructing locally aligned prognostication models without full retraining, thereby improving applicability across heterogeneous treatment environments.

Time-dependent interpretability analysis based on the permutation-based feature importance revealed interesting time-varying prognostic contributions. Age exhibited greater prognostic importance in the external GBM-only cohort than in the internal dataset (Fig. 3a), likely reflecting its established role as a key clinical factor in GBM. This observation is consistent with previous study identifying age as an independent prognostic factor in GBM patients[31]. EOR showed peak prognostic importance in early period (i.e., postoperative 3–12 months), which is consistent with previous study[36,37], while the prognostic impact of MGMT methylation was most prominent after 18 months, reflecting its role in early treatment response, especially in relation to surgical tumor reduction and sensitivity to chemotherapy[38,39] As shown in Fig. 3b, this delayed importance was particularly evident in the internal cohort, where MGMT's contribution increased over time, likely due to the prognostic role of chemosensitivity in IDH-mutant diffuse gliomas[30]. In contrast, in the external GBM-only cohort, MGMT significance plateaued or slightly decreased over 12 months, consistent with previous findings that long-term predictive value may be attenuated in this subgroup[40,41]. Supplementary Fig. 5 shows a steady increase in the contribution of image features as the number of few-shot samples increases, suggesting that few-shot fine-tuning enabled effective domain adaptation of image embeddings to the external GBM cohort. This is highly valuable for glioblastoma clinical trials of novel treatment such as IDH-inhibitor, or vorasidenib, where robust and interpretable prognostication is urgently required[42].

In the patient-specific interpretation, Grad-CAM visualizations revealed that the model's attention aligned well with annotated tumor regions, supporting the biological plausibility and interpretability of the framework (Fig. 3). Importantly, the individualized survival curves demonstrated the model's ability to differentiate between patients with distinct clinical and molecular risk profiles. In Fig. 3, the lower predicted survival probability in a patient with IDH-wildtype glioblastoma and subtotal resection reflects known poor prognostic factors, while a more favorable trajectory was observed in a patient with IDH-mutant astrocytoma and gross total resection. These findings suggest that the model not only captures spatially relevant tumor characteristics but also integrates multimodal features in a clinically coherent manner, highlighting its potential utility in personalized survival estimation and risk stratification.

Interestingly, deep survival models utilizing multiple modalities showed better calibrated risk scores, compared to those using unimodal (i.e., image-only or clinical-only) features, a finding consistent with previous study[10]. This calibration appeared to improve progressively as more data (i.e., shots) were utilized in the few-shot adaptation process (Fig. 5a). These observations suggest that multimodal deep learning, particularly when enhanced with few-shot adaptation, may offer a strategy to mitigate the overconfidence often seen in survival prediction models—a widely recognized challenge in the field[43]. The clear distinction between the risk groups in Fig. 5b reflects the model's ability to assign clinically meaningful risk scores that correspond well with actual survival outcomes.

Despite the promising performance and generalizability demonstrated in our study, several limitations should be acknowledged. First, while we curated and utilized one of the largest combined public and private datasets available for diffuse glioma, the overall sample size is still relatively limited when compared to the scale of datasets typically used in foundation model development. Nevertheless, we maximized the inclusion of all available survival-labeled data to build a robust and clinically meaningful benchmark for survival prediction. Expanding this approach to larger and more heterogeneous multi-institutional cohorts in future work could further improve the model's performance and external generalization. Second, although our use of a self-supervised learning strategy (DINOv2) helped minimize the need for manual annotations, the image encoder was still trained exclusively on glioma cases with survival labels. This introduces a risk of bias toward specific tasks or disease subtypes. To develop a more generalizable foundation model for neuro-oncology, future efforts should explore pretraining on a broader range of unlabeled brain MRI images—including healthy subjects, other neuropathologies, and various imaging protocols—to capture task-agnostic and domain-invariant representations With continued expansion of diverse training data, we expect the proposed framework to evolve into a foundation model for survival prediction in neuro-oncology. Third, the imputation of missing clinical variables using cohort-specific mean or mode values may introduce bias by smoothing clinically meaningful variability. Although this is a common strategy in large-scale clinical modeling, more advanced imputation approaches may further reduce this limitation. Fourth, our external validation relied on a single dataset composed exclusively of glioblastoma, which constrains the extent of external generalizability that can be demonstrated. Additional multi-institutional and multi-subtype validations will be necessary as more benchmark datasets become available.

In conclusion, our few-shot multimodal framework demonstrates that self-supervised image embeddings can be adapted to produce accurate, interpretable survival predictions in glioma using limited labeled data. By addressing longstanding challenges in image–clinical alignment, generalizability, and interpretability, this study contributes to the development of scalable and trustworthy AI solutions for neuro-oncology.

Data availability

The in-house datasets are not publicly accessible due to hospital regulations to protect patient privacy. Restricted access may be granted on reasonable request by contacting the corresponding author.

Code Availability

The code associated in this study will be made publicly available via GitHub: https://github.com/ggomaeng514/DeepMSM.

Received: 28 October 2025; Accepted: 24 December 2025
Published online: 14 January 2026

References

1. Reuss, D. E. et al. Heterogeneity of DNA methylation profiles and copy number alterations in 10782 adult-type glioblastomas, IDH-wildtype. *Free Neuropathol.* **5** (7). https://doi.org/10.17879/freeneuropathology-2024-5345 (2024).
2. Lucas, C. G. et al. Longitudinal multimodal profiling of IDH-wildtype glioblastoma reveals the molecular evolution and cellular phenotypes underlying prognostically different treatment responses. *Neuro Oncol.* **27**, 89–105. https://doi.org/10.1093/neuonc/noae214 (2025).
3. Hu, L. S. et al. Integrated molecular and multiparametric MRI mapping of high-grade glioma identifies regional biologic signatures. *Nat. Commun.* **14**, 6066. https://doi.org/10.1038/s41467-023-41559-1 (2023).
4. Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R. & Bridge, C. P. A review of deep learning for brain tumor analysis in MRI. *NPJ Precis Oncol.* **9** https://doi.org/10.1038/s41698-024-00789-2 (2025). 2.
5. Lee, J. O. et al. Added prognostic value of 3D deep learning-derived features from preoperative MRI for adult-type diffuse gliomas. *Neuro-Oncology* **26**, 571–580. https://doi.org/10.1093/neuonc/noad202 (2023).
6. Niu, W. et al. MRI transformer deep learning and radiomics for predicting IDH wild type TERT promoter mutant gliomas. *Npj Precision Oncol.* **9** https://doi.org/10.1038/s41698-025-00884-y (2025).
7. van der Voort, S. R. et al. Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. *Neuro-Oncology* **25**, 279–289. https://doi.org/10.1093/neuonc/noac166 (2022).
8. Wu, X. et al. Biologically interpretable multi-task deep learning pipeline predicts molecular alterations, grade, and prognosis in glioma patients. *Npj Precision Oncol.* **8**, 181. https://doi.org/10.1038/s41698-024-00670-2 (2024).
9. Gomaa, A. et al. Comprehensive multimodal deep learning survival prediction enabled by a transformer architecture: A multicenter study in glioblastoma. *Neuro-Oncology Adv.* **6** https://doi.org/10.1093/noajnl/vdae122 (2024).
10. Mahootiha, M. et al. Multimodal deep learning improves recurrence risk prediction in pediatric low-grade gliomas. *Neuro-Oncology* **27**, 277–290. https://doi.org/10.1093/neuonc/noae173 (2024).
11. Steyaert, S. et al. Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Commun. Med.* **3**, 44. https://doi.org/10.1038/s43856-023-00276-y (2023).
12. Yuan, Y. et al. Multimodal data integration using deep learning predicts overall survival of patients with glioma. *View* **5**, 20240001 (2024).
13. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
14. Steyaert, S. et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat. Mach. Intell.* **5**, 351–362 (2023).
15. Liu, Z. et al. A review of self-supervised, generative, and few-shot deep learning methods for data-limited magnetic resonance imaging segmentation. *NMR Biomed.* **37**, e5143 (2024).
16. Kirillov, A. et al. Segment anything. pp. 4015–4026. (2023).
17. Zhu, J., Hamdi, A., Qi, Y., Jin, Y. & Wu, J. Medical Sam 2: segment medical images as video via segment anything model 2. (2024). arXiv preprint arXiv:2408.00874.

18. Liu, Z. et al. vision pp. 10012–10022. (2021).
19. Kondepudi, A. et al. Foundation models for fast, label-free detection of glioma infiltration. *Nature* **637**, 439–445. https://doi.org/10.1038/s41586-024-08169-3 (2025).
20. Li, T. et al. CancerGPT for few shot drug pair synergy prediction using large pretrained Language models. *Npj Digit. Med.* **7**, 40. https://doi.org/10.1038/s41746-024-01024-9 (2024).
21. Xia, Y. et al. CT-based multimodal deep learning for non-invasive overall survival prediction in advanced hepatocellular carcinoma patients treated with immunotherapy. *Insights into Imaging.* **15**, 214 (2024).
22. Sun, Z., Li, X., Liang, H., Shi, Z. & Ren, H. A deep learning model combining multimodal factors to predict the overall survival of transarterial chemoembolization. *J. Hepatocellular Carcinoma*, 385–397. (2024).
23. Khader, F. et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Sci. Rep.* **13**, 10666 (2023).
24. Oquab, M. et al. Dinov2: learning robust visual features without supervision. (2023). arXiv preprint arXiv:2304.07193.
25. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. (2020). arXiv preprint arXiv:2010.11929.
26. Lee, D., Park, H. & Lee, C. Toward a Well-Calibrated discrimination via survival Outcome-Aware contrastive learning. *Adv. Neural. Inf. Process. Syst.* **37**, 30985–31014 (2024).
27. Cox, D. R. Regression models and life-tables. *J. Roy. Stat. Soc.: Ser. B (Methodol.)*. **34**, 187–202 (1972).
28. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. (2008). Random survival forests.
29. Gerds, T. A., Kattan, M. W., Schumacher, M. & Yu, C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat. Med.* **32**, 2173–2184 (2013).
30. Mogensen, U. B., Ishwaran, H. & Gerds, T. A. Evaluating random forests for survival analysis using prediction error curves. *J. Stat. Softw.* **50**, 1–23 (2012).
31. Kotula, C. A. et al. Comparison of multimodal deep learning approaches for predicting clinical deterioration in ward patients: observational cohort study. *J. Med. Internet. Res.* **27**, e75340 (2025).
32. Kilim, O. et al. Physical imaging parameter variation drives domain shift. *Sci. Rep.* **12**, 21302. https://doi.org/10.1038/s41598-022-23990-4 (2022).
33. Pooch, E. H., Ballester, P. & Barros, R. C. Can we trust deep learning based diagnosis? The impact of domain shift in chest radiograph classification. (Springer), pp. 74–83. (2020).
34. Silva-Rodriguez, J., Hajimiri, S., Ben Ayed, I. & Dolz, J. A closer look at the few-shot adaptation of large vision-language models. pp. 23681–23690. (2024).
35. Vettoruzzo, A., Bouguelia, M. R., Vanschoren, J., Rögnvaldsson, T. & Santosh, K. Advances and challenges in meta-learning: A technical review. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 4763–4779 (2024).
36. Karschnia, P. et al. Prognostic evaluation of re-resection for recurrent glioblastoma using the novel RANO classification for extent of resection: A report of the RANO resect group. *Neuro-Oncology* **25**, 1672–1685. https://doi.org/10.1093/neuonc/noad074 (2023).
37. Karschnia, P. et al. The oncological role of resection in newly diagnosed diffuse adult-type glioma defined by the WHO 2021 classification: a review by the RANO resect group. *Lancet Oncol.* **25**, e404–e419 (2024).
38. Balana, C. et al. A phase II randomized, multicenter, open-label trial of continuing adjuvant Temozolomide beyond 6 cycles in patients with glioblastoma (GEINO 14 – 01). *Neuro-Oncology* **22**, 1851–1861. https://doi.org/10.1093/neuonc/noaa107 (2020).
39. Wen, P. Y. et al. Glioblastoma in adults: a society for Neuro-Oncology (SNO) and European society of Neuro-Oncology (EANO) consensus review on current management and future directions. *Neuro Oncol.* **22**, 1073–1113. https://doi.org/10.1093/neuonc/noaa106 (2020).
40. Szylberg, M. et al. MGMT promoter methylation as a prognostic factor in primary glioblastoma: A Single-Institution observational study. *Biomedicines* **10**, 103390biomedicines10082030 (2022).
41. Cao, V. T. et al. The correlation and prognostic significance of MGMT promoter methylation and MGMT protein in glioblastomas. *Neurosurgery* **65**, 866–875. https://doi.org/10.1227/01.NEU.0000357325.90347.A1 (2009). discussion 875.
42. Mellinghoff, I. K. et al. Vorasidenib in IDH1-or IDH2-mutant low-grade glioma. *N. Engl. J. Med.* **389**, 589–601 (2023).
43. Kamran, F. & Wiens, J. Estimating calibrated individualized survival curves with deep learning. In **1**. pp. 240–248. (2021).

Author contributions

M.H. and J.L. designed the study. M.H., J.L., and C.L. performed data preprocessing, model development, and statistical analysis. M.K. and S.S.A. curated clinical data and contributed to manuscript drafting. S.H.C. and K.S.C. provided clinical supervision and revised the manuscript. C.L. provided technical oversight and contributed to model implementation and performance evaluation. C.L. and K.S.C. acquired funding and provided overall project supervision. All authors participated in data interpretation, manuscript editing, and approved the final version of the manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00251022) (K.S.C.); the Phase III (Postdoctoral fellowship) grant of the SPST (SNU-SNUH Physician Scientist Training) Program (K.S.C.); the SNUH Research Fund (No. 04-2024-0600) (K.S.C.); the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant funded by the Ministry of Health&Welfare (No. RS-2024-00439549) (K.S.C.); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079) Artificial Intelligence Graduate School Program (Korea University) (C.L.); and the Artificial Intelligence Star Fellowship Support Program to nurture the best talents (IITP-2025-RS-2025-02304828) grant funded by the Korea government (MSIT) (C.L.).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

Data collection was approved by the Institutional Review Boards of Seoul National University Hospital (SNUH) and Severance Hospital, with a waiver of informed consent. All procedures followed the principles of the Declaration of Helsinki. Analyses of the de-identified public datasets (UCSF-PDGM and UPenn-GBM) were exempt from IRB approval under institutional policies for secondary use of anonymized data.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-34134-9.

Correspondence and requests for materials should be addressed to C.L. or K.S.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.