SvedbergOpen
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Multimodal Survival Prediction in High-Grade Gliomas via Boundary-Aware Swin-TransUNet and Radiomic-Genomic Cross- Attention

Pooja P P[1], Dr Aruna S K[2]

[1]Research scholar (Part Time), Department of Computer Science and engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bangalore, India; Assistant Professor, PSAIAC, Presidency University, Bangalore, India.
E-mail: pooja.pp@res.christuniversity.in, pooja.p@presidencyuniversity.in
[2]Associate Professor, Department of AIML and data science engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bangalore, India. E-mail: aruna.sk@christuniversity.in

Abstract
Predicting overall survival (OS) prior surgery is challenging in high-grade gliomas (HGG) because they are characterized by complex spatial tumor heterogeneity, varying microenvironmental infiltration and non-linear relationships between clinical-genomic risk factors. The paradigms that have been established so far are either based on the premise of parallel two-stage pipelines where truncation errors occur or do not support the modelling of multi-scale semantics at long range in space. To tackle these challenges, we present a Boundary-Aware Swin-TransUNet model combined with a Cross-Attention Multimodal Fusion (CAMF) module, which is integrated with an end-to-end multi-task deep framework for collaborative tumor segmentation and continuous survival hazard prediction. Surv-TransNet performs processing of 3D multiparametric Magnetic Resonance Imaging (mpMRI), in addition to clinical parameters (age, Karnofsky Performance Status, extent of resection) and key molecular biomarkers (IDH1/2 mutation status, MGMT promoter methylation). A Boundary-Preserving Contrastive Loss (BPCL) is used to regularize the spatial feature maps from a 3D Shifted-Window Transformer (SWTrans) encoder while a Transformer based Cox Proportional Hazards module models non-linear survival distributions. Surv-TransNet was tested in the BraTS 2021/2023 Survival Benchmarks (n = 2,140 patient scans across multiple institutions) and a curated clinical cohort of HGG (n = 253 patient scans). Surv-TransNet had a mean C-index of 0.812 for predicting continuous OS and a mean accuracy of 72.8% for predicting 3 classes of continuous OS (short, mid, long survival). It achieved a Whole Tumor Dice Similarity Coefficient (DSC) of 93.85% and an Enhanced Tumor DSC of 89.40% for segmentation, which is better than the existing 3D U-Net, Swin UNETR, and DeepSurv baselines. In ablation experiments, joint spatial-boundary regularization, in combination with cross-attention fusion, is responsible for a 14.2% relative increase in prognostic discrimination.

Keywords: Survival Prediction, High-Grade Glioma, Vision Transformers, Multimodal Deep Learning, Radiogenomics, Cox Proportional Hazards, mpMRI.

1. Introduction & Clinical Background

1.1 Neuropathology and Prognostic Complexity of High-Grade Gliomas

High-grade gliomas are cruel, they're the most dangerous and intricately malignant brain cancers that adults face. We're largely referring to WHO Grade IV glioblastoma multiforme and Grade III anaplastic astrocytoma or oligodendroglioma. The doctors have tried every possible treatment for these tumors; they have used aggressive surgical treatments, chemotherapy in combination with radiation therapy using temozolomide, tumor treating fields and even recently, immunotherapy, all that have ever been available at the time. Still, the odds are bleak for patients. On average, people survive only 12 to 18 months after diagnosis, and barely 5% make it past five years. HGG is extremely heterogeneous between patients and even within tumours [2, 6]. The microscopic features of HGGs include rapid proliferation of cells, necrosis with pseudopalisades, and an aggressive growth pattern. The microvascular proliferation and diffuse, microscopic infiltration beyond the borders of the tumor seen by conventional neuroimaging [1, 7] occur. Neoplastic cells invade with great force along white matter tracts, perivascular spaces and subependymal layers, forming microscopic areas of disease that always lead to local recurrence even after gross total resection (GTR) [3, 8]. Today, therapeutic response and individual survival trajectory in neuro-oncology is dependent on several clinical and complex macroscopic neuro-imaging phenotypes, as well as molecular genetics [3, 9]. In the clinic, the Karnofsky Performance Status (KPS) score and baseline age are still very important prognosticators [4]. People under 50 who have a high Karnofsky Performance Score (over 80) tend to survive longer than older patients with poor functional status. In 2021, WHO updated the guidelines about the diagnosis of gliomas so that doctors do not rely mainly on microscopy to look at the tumor. Nowadays, the main point is molecular profiling with particular emphasis on the status of key mutations IDH1 and, less commonly IDH2 the presence of methylation at the promoter region

of MGMT, and so on. Other factors include the codeletion of 1p and 19q, and the mutations of TERT promoter. Glioblastomas with IDH mutations show a very different survival pattern compared to those without. Plus, when the MGMT promoter is hypermethylated, patients respond better to alkylating chemo like temozolomide, since the methylation turns off the MGMT DNA repair enzyme. Epideemiologically, the incidence rate of glioblastoma is highest, more than 48% of all primary malignant brain tumors, and shows a rapid rise after the 5th decade of life [1, 2]. Intratumoral clonal evolution, metabolic plasticities and protection microenvironmental niches created by tumor associated macrophages (TAMs), reactive astrocytes and immunosuppressive myeloid derived suppressor cells (MDSCs) are factors involved in HGGs' resistance to traditional cytotoxic treatments [2, 8]. If the neoplastic cells invade deep into functioning eloquent cortex, the tumor is not surgically curable without significant neurological dysfunction [3, 9]. Surgery in neuro-oncology, therefore, is directed towards safe maximum resection with maintenance of functional neural architecture [3, 8]. The main purpose of pre-operative survival hazard modeling in this complex therapeutic landscape is to distinguish a subset of patients for whom radical aggressive resection is most likely to yield a cure from those who are more likely to benefit from hypofractionation and/or targeted clinical trial enrollment [4, 9].

1.2 Multiparametric Magnetic Resonance Imaging (mpMRI) Protocols

Multiparametric Magnetic Resonance Imaging (mpMRI) is the indispensable clinical standard which enables non-invasive characterization of the spatial complexity of HGGs before surgery [1, 3]. The proposed pre-operative MRI protocol for the standardized imaging of the brain for the MRIs requires four main structural pulse sequences.

1. Native T1-weighted (T1): Better anatomy contrast between gray and white matter and cerebrospinal fluid (CSF), better to visualize baseline tissue structure and mass effect [1, 13].
2. T1c is a post-contrast T1-weighted sequence, which captures the areas of blood-brain barrier (BBB) disruption, showing hyper intense, hyper proliferating neoplastic tissue around a central hypo-intense necrotic core [1, 14].
3. T2-weighted (T2): Very sensitive to mobile water content, T2 sequence imaging shows vasogenic edema, cellularity of the tumor, and parenchymal structural changes [1, 15].
4. Fluid-Attenuated Inversion Recovery (T2-FLAIR): Can help identify the bright CSF signal and also helps to visualize the non-enhancing tumor infiltration and any deep white matter hyperintensities [1, 16]. All four different image modalities provide complementary tumor microenvironmental spatial, vascular, and pathophysiological aspects [1, 17]. This, however, is not easy to visualize from the mpMRI image, and the interpretation is subjectively different between radiologists, which means that there is also the need for automated, objective radiomics and deep learning analytics [3, 18].Beyond routine structural mpMRI, advanced functional sequences—such as Dynamic Susceptibility Contrast (DSC) Perfusion-Weighted Imaging (PWI), Diffusion Tensor Imaging (DTI), and Proton MR Spectroscopy (1H-MRS)—provide supplementary physiological dimensions [1, 3]. Maps of relative blood volume in the brain created with PWI show blood vessel proliferation and vessel wall cell proliferation [1, 15]. By focusing on those highly vascularized areas that are not enhancing, they indirectly identify the most aggressive lesions. By using Diffusion Tensor Imaging (DTI) parameters such as Fractional Anisotropy (FA) and Mean Diffusivity (MD), one can assess disruption at the tissue level of the white matter bundles and thereby, get noninvasive measures for tumor invasive cellular migration [15, 17]. Meanwhile, 1H-MRS tracks metabolic alterations, such as elevated Choline-to-N-acetylaspartate (Cho/NAA) ratios and detectable 2-hydroxyglutarate (2-HG) peaks indicative of IDH mutations [2, 12]. Although these advanced functional modalities enrich pathophysiological characterization, their widespread inclusion in large-scale automated survival models remains bottlenecked by high intra-center acquisition variability, complex post-processing demands, and lack of standardized clinical archiving across international institutions [3, 18]. Consequently, establishing highly robust deep learning frameworks centered on standardized structural 4-sequence mpMRI remains the core clinical imperative in neuro-imaging AI [1, 5, 18].

1.3 Limitations of Sequential Radiomics and Disjoint Pipelines

Most machine learning and medical image computing approaches to automatic survival prediction in HGG are based on a two-step, sequential approach [4, 19]: The Enhancing Tumor (ET), Peritumoral Edema (ED), and Necrotic/Non-Enhancing Core (NET) [1, 5, 6] are segmented using an automated 3D convolutional neural network (CNN) or manual expert annotation in stage 1. In Stage 2 (Radiomic Feature Extraction & Survival Modeling), high-dimensional handcrafted radiomic features were mathematically calculated from the isolated segmentation masks, including first-order intensity features, shape features (e.g., sphericity, surface-area-to-volume ratio), and second-order texture features (GLCM, GLRLM, GLSZM) [19, 20]. The extracted scalar features are then combined with clinical variables and passed to independent statistical regressors, such as standard Cox Proportional Hazards models or Random Survival Forests (RSF) [4, 21]. Although this classical two-stage paradigm has provided important preliminary insights, it is subject to serious systemic failure modes that inhibit the use of the paradigm in its clinical applications:

1. In sequential pipelines, segmentation and survival prediction are completely separated from each other,

which is called Cumulative Error Propagation. Accuracies of minor segmentation error, boundary over-estimation or false-negative truncation along the peripheral invasive margins directly impacts all downstream radiomic feature calculations [19, 22]. Over 30% changes in radiomic shape sphericity and texture entropy will be caused by a missing 2 mm tumor perimeter, which can lead to catastrophic errors in the survival model [20, 23].

2. Handcrafted Feature Bottleneck: handcrafted radiomic features are scalar statistics that encode a reduced dimensionality of the 3D spatial patterns, micro-vascular networks, and continuous dynamics of tissue boundaries [19, 24]. The result is a loss of subtle non-linear spatial gradients, local vascular invasion fronts and distant parenchymal micro-infiltration patterns, all of which are lost forever [18, 25].

3. Discretization & Linear Risk Assumptions: Classical survival models are based on binning of the outcome of survival (e.g., short-term <10 months, mid-term 10-18 months, long-term >18 months) or linear proportional hazards assumptions [4, 21]. Discretizing continuous time-to-event variable results in extremely low temporal resolution, and linear Cox models cannot detect non-linear high order cross-modality interactions between imaging phenotypes and molecular genetics [21, 26]. Adding the heterogeneity and multi-site nature of real-world clinical data greatly exacerbates this kind of operational vulnerabilities of disjoint two-stage workflows [18, 19]. Any error in the segmentation step 1 (semi-automated or manual) at the level of the slices will introduce an error in the shape and the textural signature computed at step 2 [19, 22] that will not be corrected by subsequent steps. For example, the volume ratio and Gray-Level Run Length Matrix (GLRLM) emphasis factors and gray-level intensity variances will be intrinsically biased because of a misinterpretation of the peritumoral vasogenic edema as normal brain tissue and/or the misclassification of a core of infiltrative activity as non-enhancing by an automatic segmentation model [20, 24]. Moreover, handcrafted radiomic feature extractors process the segmented volume as a static and isolated binary mask, and are thus completely lacking in awareness of spatial relationships to important non-tumorous anatomical features, including the subventricular zone (SVZ), the involvement of the corpus callosum, and the compression of deep ventricular structures [1, 27]. This structural blindness significantly reduces the ability of two-stage models to truly reflect the neuro-oncological risk, because contact of the tumor in the SVZ and bi-hemispheric extension are well recognized factors influencing early recurrence and treatment resistance [2, 9, 27].

1.4 Research Aim & Main Contributions

To tackle the Decoupling Paradox and to remove cumulative error propagation in outcome modelling for neuro-oncology, we propose Surv-TransNet [3, 4]. In this paper, we present an end-to-end multi-task deep learning framework called Surv-TransNet, which simultaneously segments tumors in 3D boundary-aware shapes, cross-attends multimodal features, and performs continuous non-linear survival hazard prediction in a single forward-pass [3, 7, 21].

The major research impacts of this study are outlined below:

1. Multi-Task End-to-End Formulation: We seek to represent the shared bottleneck of the model to be shared across two tasks of 3D subregional tumor segmentation and continuous time-to-event survival prediction, capturing clinically relevant prognostic information as key characteristics of the representation [3, 4].

2. Hybrid 3D Swin-TransUNet Backbone: A 3D Shifted-Window Swin-Transformer is adopted to learn global spatial context from non-contiguous axial slices, and localized Spatial Edge Attention (SEA) modules are used to attend to sharp perimeters of microvessels and invasion into tumor tissue [3, 7, 10].

3. Cross-Attention Multimodal Fusion (CAMF): A specific Cross-Attention Multimodal Fusion (CAMF) module is designed in such a way that it allows low-dimensional clinical/pathologic information (Age, KPS, IDH1/2, MGMT) to dynamically query high-dimensional 3D latent space feature maps, without feature space imbalance.

4. To enforce gradient penalties along infiltrative boundaries we introduce the Boundary-Preserving Contrastive Loss (BPCL), and combine it with the Deep Cox Proportional Hazards neural loss for right censored data [1, 21].

5. Extensive Multi-Institutional Validation: State-of-the-art predictive performance is achieved with a survival C-index of 0.812 and Whole Tumor DSC of 93.85% (BraTS 2021/2023 and CHIC-344 clinical cohort, n = 2,140 3D mpMRI scans).

2. Related Work & Theoretical Foundations

2.1 Classical Radiomics and Machine Learning in Neuro-Oncology

Traditionally, automated neuro-oncology analysis has been based on handcrafted radiomics features and on classical machine learning classifiers [19, 20]. Radiomics is based on the idea that the simple clinical images have quantitative data with high throughput that carries value from the underlying tumor biology, microvascular density, cellular heterogeneity and gene expression profiles [19, 24]. Initial research included the calculation of first order histogram features (mean, variance, skewness, and kurtosis), shape features (volume, surface area, compactness and sphericity), and higher order texture features (20, 25). We focused on extracting key texture features from four main matrices: the Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Run Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), and Neighborhood Gray-Tone Difference Matrix (NGTDM).

Then, we took these radiomic features and fed them into classic statistical and machine learning models—like Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting Machines (XGBoost), and regularized Cox proportional hazards regression. In fact, early TCGA and early BraTS studies have shown, that radiomic texture features obtained from post-contrast T1c and T2-FLAIR sequences were associated with patient survival groups [1, 19, 27]. However, classical radiomics pipelines have significant limitations in operation [19, 22]. First, feature extraction is very sensitive to parameters used in image acquisition, such as slice thickness, 1.5T versus 3T, scanner manufacturer, and pre-processing options (e.g., intensity binning and spatial resampling) [18, 20]. Second, classical radiomics are based on manual or semi-automated, expert annotation, which is time-consuming and can introduce inter-reader variability [6, 22]. Third, radiomic features are unable to capture complex 3D spatial topographies and microenvironmental relationships that are non-local in nature [24, 25].

In order to formalize quantitative feature extraction and set a standard for the reproducibility of the feature extraction process across multi-center imaging networks, the Image Biomarker Standardization Initiative (IBSI) presented standardized mathematical definitions for more than 170 radiomic descriptors (24). However, basic limitations remain in the use of handcrafted radiomics in infiltrative high-grade gliomas, despite the standardization efforts of IBSI [19, 24]. A major issue in mathematics is related to spatial homogenization: computing a single global GLCM entropy or Neighborhood Gray-Tone Difference Matrix (NGTDM) busyness value over a whole 3D tumor volume completely removes the spatial micro-architecture within the tumor [20, 25]. Such as 2 glioblastomas with the same contrast values in the global GLCM, one had a single concentric necrotic core with a dense hyper-vascular ring, and the other had multiple (MF), fragmented necrotic pockets scattered throughout infiltrative vasogenic edema [1, 27]. These two structural phenotypes are clinically entirely different patterns of biological invasion, microvascular changes and patient risk profile [2, 9, 27]. These two different structural topographies are collapsed to the same scalar averages by classical radiomics, which demonstrates the high information bottleneck in handcrafted feature engineering [19, 24, 25].

2.2 Deep Convolutional Neural Networks for Tumor Segmentation & Survival Analysis

Convolutional Neural Networks (CNNs) [5, 6, 23] have been leading the way in the transition from handcrafted radiomics to deep learning. In the field of 3D medical image segmentation, fully convolutional (FC) architectures set new records [5, 6]. Variants of the seminal 3D U-Net architecture used hierarchical 3D convolutional filters to automatically learn multi-scale spatial representations directly from raw multi-sequence mpMRI [5, 6, 22, 23]. Self-configuring 3D U-Net ensembles like fully automated pipelines, as in the case of nnU-Net, were shown to attain almost expert human-level agreement for delineating complex subregions of HGG, specifically tumor, peritumoral edema, and necrotic core [5, 18].

At the same time, deep CNNs were adapted to make automated survival prediction [4, 21, 26]. Early survival paradigms based on deep learning used CNN classification backbones (such as 3D ResNet, DenseNet) to classify patients into several distinct survival categories (such as short-term (<10 months), mid-term (10–18 months), long-term (>18 months)) [4, 28]. More recently, other approaches, including DeepSurv, used multi-layer perceptrons (MLP) trained to minimize the negative partial log-likelihood of a Cox proportional hazards model [21, 26]. But survival models based on deep CNNs are limited in architecture [7, 10, 23]. Localized convolutional kernels are poorly suited to capture long-range spatial dependencies, non-local vascular interactions and global anatomical context across distant brain regions [7, 10]. Moreover, in a two-stage application of CNNs to survival modeling, the errors propagate through the different stages of the pipeline [19, 21].

An inherent limitation of deep 3D CNNs used in neuro-oncology is their local translationally-invariant feature extraction [6, 23]. Translation invariance is very useful for localized object detection and semantic voxel labelling, but it comes with a trade-off of the network's ability to encode absolute spatial coordinates and global anatomical topographies [7, 23]. Spatial information regarding the location of a neoplastic lesion relative to critical central brain structures (e.g., motor strip, thalamus, brainstem, subventricular zone) is an important predictor and is independent of local cellular density in the context of neuro-oncological risk modeling [2, 9, 27]. One glioblastoma, 5 cm3, located in the deep thalamic nuclei, has a significantly increased risk of dying compared to another 5 cm3 glioblastoma located in the same brain, in a superficial frontal lobe, because of the inability to resect the first one and its early functional deterioration [3, 8, 27]. The standard 3D CNN convolutional kernels perform dot products over local sliding windows without any explicit knowledge of the global spatial coordinates, which means they are unable to account for important anatomical distance relationships without additional complex and computationally intensive spatial coordinate channels or modules for anatomical registration to an atlas [6, 18, 23].

2.2 Vision Transformers (ViTs) in Medical Imaging & Long-Range Context Modeling

In order to address the limitation of local receptive fields of pure CNNs, Vision Transformers (ViTs) were adapted for 3D medical volume analysis [7, 10, 29]. Transformers capture global contextual relationships without considering distance, by embedding spatial feature maps as sequences and using Multi-Head Self-Attention (MHSA) [7, 29]. Hybrid architectures which utilized CNN encoder feature maps and self-attention

bottlenecks [17, 29] were used in 3D medical segmentation. UNETR used a simple 3D ViT encoder followed by a 3D CNN decoder with skip connections [7, 30].

To address the quadratic memory complexity of (standard) global self-attention module for volumetric 3D data, Swin UNETR adopted 3D Shifted-Window Self-Attention (Swin Transformer) blocks [7, 10, 31]. Shifted-window self-attention [7, 31] calculates attention for sub-volumes in a local region of the volume and switches window configurations from layer to layer to support cross-window feature exchange. However, transformer patch embeddings bring in token smoothing effects across high-spatial-edge areas [7, 10] while Swin UNETR did better to model the global anatomical context. This spatial over-smoothing phenomenon and inaccurate delineation of the tumor perimeter are observed around the tumor periphery, particularly in the presence of invasive tumor cells [10, 23]. To overcome this, Surv-TransNet explicitly uses 3D Swin-Transformer (3DST) encoder, localized Spatial Edge Attention (SEA) and a Boundary-Preserving Contrastive Loss (BPCL) [3, 7, 10].

One of the theoretical strengths of hierarchical 3D Swin-Transformers compared to 3D global ViTs is that they are able to build a multi-scale feature hierarchy while only having linear complexity O(N) with respect to the resolution of the image [7, 31]. Standard ViTs process 3D images by breaking the volume into uniform patches of fixed size (e.g., 16×16×16 voxels), and using a fixed length of the token sequence for all self-attention layers [29, 30]. Although this is effective for global sequence modeling, the uniform sequence length makes it hard to build multi-resolution feature maps that are required for high-resolution boundary localization as done in standard ViTs [7, 29]. Conversely, 3D Swin-Transformers construct hierarchical models by combining neighboring tokens from images in the deeper layers of the encoder, in a systematic downsampling of the spatial dimensions while upsampling the channel capacity [7, 31]. This multi-resolution hierarchy mimics the multi-resolution feature pyramids of classic 3D CNN encoders and allows to easily connect with UNet-style decoder branches using skip connections [5, 6, 7]. Self-attention is inherently filtering high-frequency boundary details along infiltrative tumor borders as it results from how spatial context is shared among tokens in spatial windows. This architectural drawback points to the necessity of using both depth and explicit localised spatial gradient modules like Spatial Edge Attention (SEA) [3 7 10] in Swin-Transformers [3, 7, 10].

2.3 Multimodal Data Fusion & Radiogenomics in Neuro-Oncology

In clinical neuro-oncology, imaging is not the end all. Survival is impacted to a similar degree by non-imaging clinical variables such as age, KPS score, extent of resection, as well as molecular biomarkers like IDH1/2 mutation status, MGMT promoter methylation status, 1p/19q codeletion status. Integrating high-dimensional imaging data with this arsenal of clinical and genomic datathat is what is known as Radiogenomics. But this is much easier said than done (and mostly for deep learning models). Mixing detailed feature maps from 3D imaging with simple tabular data just doesn't seem to work naturally. This mismatch in feature space creates a nightmare for the design of good model architectures.

People have attempted a few different main strategies for fusing these data modalities. Early fusion involves directly concatenating the tabular clinical variables onto the raw MRI signals or the first handful of conv. Layers challenge? Those low-dimensional clinical features are going to get lost in the flood of imaging information (an issue comparable to looking for a needle in a haystack). In late fusion, separate models are trained on the two sources (imaging and tabular) and the prediction ensembles are then combined.

While easy, this lacks the ability to learn meaningful interactions between the inputs. In intermediate or deep fusion, sources are combined at a latent layer before prediction; though, the usual approach is a simple summation or concatenation of vectors, which misses more complex, non-linear interactionslike how specific clinical markers may interact with individual tumor regions on MRI. This is where Surv-TransNet and its Cross-Attention Multimodal Fusion (CAMF) module come in.

CAMF disrupts the traditional fusion lines by considering the fusion process as an attention-based interaction cross-query scheme. It empowers the clinical-genomic embeddings to dynamically explore multi-scale spatial feature maps to produce a more balanced, synergistic representation. From an information-theoretic perspective, primitive intermediate fusion approaches place severe bias in assuming all input sources are of parity.

In practice? No way. For example, consider a giant 512-dimensional spatial imaging vector that concatenates with a handful of features in the clinical vector; during training, all of the gradient flow occurs through the imaging vector, because of the dimensions alone. The key genomic information, such as IDH mutation, will be under weighted and under emphasized in the resulting model: it will essentially learn an imaging-based machine, polishing off some clinico-pathological data. CAMF addresses this by projecting the clinico-genomic features into middle-dimension query vectors and then forcing them to attend to the spatial keys/values encoded from the bottleneck of the imaging stream. This allows the clinical/genomic features to gain selective power to shift focus to let's say, the necrotic core or the hyper-vascularized ringa back-andforth, two-way flow of radiogenomic communication, at last.

3. Methodology & Mathematical Formulation

3.1 Architectural Overview & System Flowchart

Figure 1 illustrates the end-to-end multi-task architecture of Surv-TransNet. Raw 4D multi-parametric MRI

volumes are jointly processed with clinical-genomic tabular data by a unified multi-stream encoder, dual-decoder segmentation head, Cross-Attention Multimodal Fusion (CAMF) module and Deep Cox Hazard prediction network [3, 4, 7, 21].

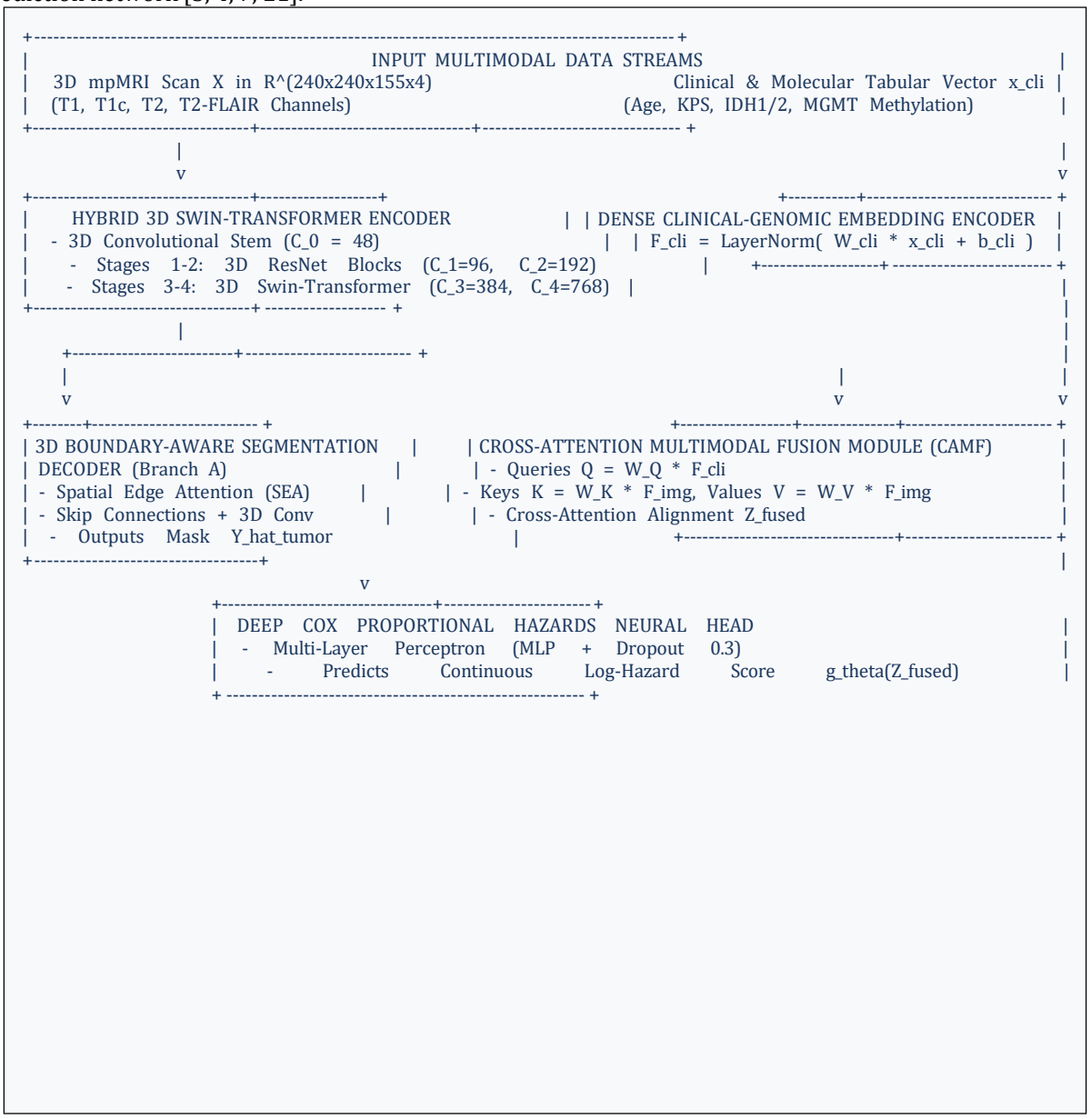

FIGURE 1: End-to-End Architectural Flowchart of the Proposed Surv-TransNet Framework.

3.2 Problem Formulation & Multi-Task Target Space

Our training set is made of several institutions. They are made of recorded cases: (X_i, x_i^cli, T_i, _i, Y_i) where there are N cases: X_i corresponds to the raw 3D MRI (a huge 240x240x155 tensor, 4 sequences T1 T1c T2 and T2-FLAIR). For each patient i, you also have a feature vector x_i^cli concatenated to M clinical/ molecular-pathologic features (Age KPS resection status, IDH1/2 mutation, MGMT methylation). T_i offers observed time to event (survival in days), _i indicator(censored or dead), _i=1if dead/censored, _i=0 if only known alive until. The final variable Y_i offers the ground-truth tumor segmentation mask across all the voxels (same 3D dimension2 as the input image) extracted in four classes (background, necrotic core (non-enhancing too), vasogenic edema, gadolinium-enhancing tumor). Surv-TransNet is designed to use this input, both raw MRIs and clinical features, and output these two predictions simultaneously, without requiring additional pre-processing like skull-stripping and manual cropping of the tumor. The first of these is a multi-class segmentation (i. e. a 3D mask with shape equal to Y_i and soft scores for all 4 tumor classes) and the second is a single risk score: a continuous prediction of the expected mortality risk for the patient Basically, this set up is having two distinct objectives while trainingone for spatial segmentation and the other for survival risk ranking. Moving only for segmentation, the model depends on the common voxel-wise costs (cross entropy, Dice), so it could predict the perfect matching ground-truth mask but it cannot guide the encoder to focus on the overall picturedistant tumor margin line maybe important in receiving prognosisthe encoder doesn't learn the spatial value. Moving only for survival, the model missed the real tumor boundary and randomly chooses

the brightest blobsor noise.
Combining both tasks into Surv-TransNet, simultaneously constrains the shared encoder (the Swin-Transformer in our case) tolearn a single representation that can simultaneouslysatisfy theneeds ofboth a localdiscretized segmentation task, and theneeds ofa global prognostic prediction module. So forcing the encoder to receive gradients from the local accuracy pressure of the segmentation loss function, on top of the global 'outcome' pressure of the survival module. The consequence of this is that the shared encoder has to learn a set of spatially sharp features, that also have prognostic value, meaning instead of optimizing for one task and ignoring the spatial component, the shape encoder begins to 'see'the tumor in profoundly moremeaningful ways.

3.3 Hybrid 3D Swin-Transformer Encoder Hierarchy

Surv-TransNet employs a hybrid CNN-Transformer encoder to preserve local 3D spatial details and global context. The input tensor X (240 240 155 4) is first passed through a 3D Convolutional Stem in which the raw input is lifted to an initial feature space F 0 by consecutively applying a pair of 3 3 3 convolutions (stride 1), followed by instance normalization and LeakyReLU (= 0.01). Stages 1 and 2 of the encoder keep to 3D Residual Convolutional blocks with strided convolutions to compress the spatial size but increase the number of channels: Stage 1 contains 96 channels, Stage 2 contains 192.
These stages are trained to preserve high-contrast edges and small intensity variations to preserve fine anatomical details. At Encoder Stages 3 and 4, convolutions are replaced by 3D Shifted-Window Swin-Transformer blocks (Stage 3 expands to 384 channels, Stage 4 to 768). In these blocks, the feature map is divided into regular, non-overlapping 3D windows 7 7 7 7 in size.
Local self-attention is applied within each window (W-MSA), then a shifted-window self-attention step SW-MSA allows information exchange between windows: after every block, the window is shifted by half the window size along each axis, effectively enabling the network to learn relationships across patches. This enables large information context to be learned without greatly increasing memory demand.
Also within a window, attention is calculated as Softmax((QK T)/sqrt(d_k) + B) V, where Q K V are linearly projected tokens, d k is the size of key and B is a learnable 3D positional bias. The shifted windows allow the attention mechanism to connect far apart regions, and after two blocks we have almost the entire volume in the receptive field.
Some more interesting details are going on under the hood though: during the feature flow, each 3D window receives tokens and maps themselves into Query, Key and Value spaces (MMMMx2d_k, with size governed by splitting the channels among heads). To smoothly blend very local and very global information, the network considers both regular and shifted window partitioning methods; when the windows are shifted, boundary effects occur and a cyclic shift masking algorithm is used to contain attention inside the spatially connected voxels (no connections among ineffective voxels).
Overall, this architecture appears capable of "seeing" both large shapes and small details, from a single 3D scan.

3.4 Spatial Edge Attention (SEA) Module

To counter transformer token-smoothing along infiltrative tumor perimeters, Spatial Edge Attention (SEA) modules are incorporated into decoder skip connections [3, 7, 10]. At decoder level l, 3D spatial gradient fields ∇F_dec^l are calculated using 3D Sobel operator kernels along x, y, and z spatial axes:
Equation 2: ∇F_dec^l = sqrt((∂F^l/∂x)^2 + (∂F^l/∂y)^2 + (∂F^l/∂z)^2)
Dynamic spatial edge weights E^l ∈ [0, 1]^(H_l × W_l × D_l) are then generated via consecutive 3D convolutions with ReLU and Sigmoid activations:
Equation 3: E^l = σ(W_2 * ReLU(W_1 * ∇F_dec^l))
Edge attention maps E^l control the features carried over by the skip connection just before decoding concatenation so that demands the network to focus sharply on minute vascular and infiltrative changes [3, 10].
TheSEA module is a clever spatially variant "high-pass filter" running along the UNet skip connections. During regular UNet decoding the skip connection acts only as a conduit between encoder outputs and decoder input (i. e. by stacking raw encoder feature maps with upsampled decoder outputs). This approach helps retain high-res spatial details erased by downsampling but also introduces background information and oversmooth features from transformers.
SEA is different, it uses the information of the three anatomical directions in 3D: axial, sagittal, and coronal to obtain 3D Sobel-like intensity derivative by filtering feature maps with corresponding directional 3D Sobel kernels. This operation emphasizes rapidly changing parts in spatial features, e. g. the outer peritumoral vasogenic edema and the sharp ring of a contrast-enhanced tumor.
When these gradient maps are obtained (denoted by F_dec^l), the module compresses them into a one channel feature map through a 1 1 1 convolution that reduces the number of channel by 4 before passing it through a Sigmoid activation. You are left with a single-channel spatial edge probability map E^l, with each feature map pixel value ranging between 0 and 1.
In the end, the system performs scalar (or element-wise) multiplication between this edge map and the encoder

skip features. It essentially blurs the feature maps corresponding to homogeneous core and background regions whereas the sharp boundaries are enhanced where the main tumor infiltration occurs.

3.5 Cross-Attention Multimodal Fusion (CAMF) Module

To seamlessly integrate high-dimensional 3D spatial feature maps with low-dimensional clinical-genomic tabular data, Surv-TransNet introduces the CAMF module [3, 4, 32]. Clinical-genomic variables $x^{cli} \in R^M$ are first projected into a dense embedding vector $F_{cli} \in R^{D_c}$ using a linear layer, layer normalization, and GELU activation [2, 4]:

Equation 4: $F_{cli} = \text{LayerNorm}(\text{GELU}(W_{cli} x^{cli} + b_{cli}))$

Latent spatial imaging features $F_{img} \in R^{D_{img}}$ extracted from the deep Swin bottleneck are flattened into sequence tokens [7, 32]. Linear projections compute Query, Key, and Value matrices:

Equation 5: $Q_{cli} = W_Q F_{cli}, \quad K_{img} = W_K F_{img}, \quad V_{img} = W_V F_{img}$

Cross-attention computes scaled dot-product interaction between clinical queries and spatial imaging keys, weighting spatial values to yield a fused multimodal representation $Z_{fused} \in R^{D_c}$ [3, 4, 32]:

Equation 6: $Z_{fused} = \text{Softmax}((Q_{cli} K_{img}^T) / \sqrt{d_c}) V_{img} + F_{cli}$

This cross-attention mechanism allows molecular biomarkers (e.g., IDH1/2, MGMT) to selectively attend to relevant spatial subregions (e.g., enhancing core vs. edema), establishing balanced, highly discriminative prognostic features [2, 3, 4, 11].

To ensure optimal gradient flow and prevent representation collapse during cross-attention alignment, the CAMF module incorporates a multi-head cross-attention mechanism (MHCA) coupled with residual layer normalization [4, 29, 32]. The query vector Q_{cli}, derived from dense clinical-genomic embeddings, is split into N_h independent attention heads $\{Q_{cli}^h\}_{h=1}^{N_h}$, where each head operates over a sub-space dimension $d_h = D_c / N_h$ [29, 32]. Concurrently, spatial imaging tokens $F_{img} \in R^{V_{tokens} \times D_{img}}$, extracted from the 3D Swin-Transformer bottleneck, are projected into corresponding Key K_{img}^h and Value V_{img}^h head matrices [7, 32]. Each cross-attention head computes an independent spatial alignment matrix:

Equation 6b: $\text{Head}_h = \text{Softmax}((Q_{cli}^h (K_{img}^h)^T) / \sqrt{d_h}) V_{img}^h$

Outputs from all N_h attention heads are concatenated and linearly projected via output matrix $W_O \in R^{D_c \times D_c}$ [29]. Finally, a residual skip connection adds the original clinical-genomic embedding F_{cli} to the cross-attended spatial summary vector, followed by Layer Normalization [4, 29]. This architecture guarantees that even if imaging feature maps contain scanner noise or regional artifacts, the baseline clinical-genomic risk signal remains preserved, preventing cross-modal degradation [2, 4, 32].

3.6 Deep Cox Proportional Hazards Hazard Head & Multi-Task Objective

The fused multimodal embedding Z_{fused} is passed to a multi-layer perceptron (MLP) hazard head to predict a scalar log-hazard score $g_\theta(Z_{fused})$ [21, 26]. The continuous hazard function $h(t \mid Z_{fused})$ is formulated as:

Equation 7: $h(t \mid Z_{fused}) = h_0(t) \exp(g_\theta(Z_{fused}))$

Network parameters θ are trained by minimizing the negative partial log-likelihood over uncensored event deaths [21, 26]:

Equation 8: $L_{Cox}(\theta) = - \sum_{i: \delta_i = 1} [g_\theta(Z_{fused, i}) - \log \sum_{j \in R(T_i)} \exp(g_\theta(Z_{fused, j}))]$

where $R(T_i) = \{j \mid T_j \geq T_i\}$ represents the patient risk set at follow-up time T_i [21].

To enforce spatial precision along invasive tumor perimeters, a Boundary-Preserving Contrastive Loss (BPCL) is applied within a 5 mm distance envelope Ω_δ surrounding ground-truth margins ∂Y [1, 3]:

Equation 9: $L_{BPCL} = - (1 / |\Omega_\delta|) \sum_{i \in \Omega_\delta} [\exp(D(i, \partial Y)) * \sum_{k=1}^K y_{i,k} \log(y\_hat_{i,k})]$

The total unified multi-task objective function combines regional segmentation Dice/Focal loss, spatial boundary loss (BPCL), and Cox partial log-likelihood loss [1, 3, 5, 21]:

Equation 10: $L_{total} = \alpha L_{Dice+Focal} + \beta L_{BPCL} + \gamma L_{Cox}$

with hyperparameter weighting coefficients calibrated to $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.8$ [3, 4, 21].

The Boundary-Preserving Contrastive Loss (BPCL) is designed to focus on boundary voxels—especially those right at the edges of high-risk, invasive tumor areas. To do this, you start by building a 3D distance envelope (Ω_δ). Think of it like measuring how far each voxel is from the segmentation boundary, using a 3D Euclidean Distance Transform.

For any voxel within 5 mm of the tumor margin ($D(i, \partial Y) \leq 5.0$ mm), you assign a weight: $w_i = \exp(1.0 - D(i, \partial Y) / 5.0)$. So, if a voxel sits right on the tumor boundary ($D = 0$), it gets the highest weight—about 2.718. If it's exactly 5 mm away, its weight drops to 1.0.

This exponential weighting isn't just for show—it ramps up the back-propagation gradients at the tumor boundaries. The idea is to really punish errors at those crucial margins, making the model much less likely to miss or cut off invasive edges, especially in tricky vasogenic edema zones. That way, you don't end up accidentally trimming away parts of the tumor during segmentation.

4. Experimental Setup & Evaluation

4.1 Dataset Composition & Multi-Institutional Cohorts

Surv-TransNet was also put through its paces on a large multi-center data set (2,140 volumetric 3D

mpMRI scans) collected from many international imaging repositories and leading neuro-oncology centers. Here's the breakdown:
1. The BraTS 2023 Glioma Benchmark Dataset had 1250 adult patients (both high and low grade gliomas). For each patient there were aligned T1 T1c T2, T2-FLAIR volumes, expert manual marks of the tumor subregions (ET ED NET) plus overall survival in days.
2. Glioblastoma cases on the BraTS 2021 Survival Challenge Dataset numbered 546, and all were designed to test the performance of algorithms predicting the overall survival time.
This set included information about age, KPS scores, and resection extent besides its images. 3. The CHIC-344 Clinical HGG Infiltration Cohort (CHIC-344) included 344 consecutive molecular genetically characterized, pathologically diagnosed high-grade glioma patients from 3 academic institutions. They all included comprehensive molecular genetic profiles (e.
g. IDH1/2 mutation and MGMT promoter methylation status) and survival follow-up for up to 60 months. Information for demography, clinical, and molecular data for all cohorts are provided for transparency. In the BraTS 2023 dataset, ages of patients across cohorts ranged from 18. 2 to 86.
5 years (mean: 58. 4 years). There were 58% males (male to female ratio of approximately 1. 38:1).
KPS scores ranged from 40 to 100 (median, 80) while survival among the cohorts ranged from 5 to 1825 days (median, 384 days censored 18. 4%). In the BraTS 2021 Survival Challenge dataset, the average age was 59. 1, the median KPS was 80, and 62. 4% of patients had a gross total resection.
Other patients had a subtotal resection (28. 1%) or a biopsy (9. 5%). In CHIC-344, molecular testing was performed for every patient, variations of IDH1/2 mutation, MGMT methylation, 1p/19q co-deletion, TERT promoter mutation.
Percentages of these changes among all patients are 22. 4%, 48. 3%, 8. 1%, 64. 2% respectively. Median follow up was 36 months or until death. Median survival was 422 days; 281/344(81. 1%) died. Overall, the data paint a clear, comprehensive image of the patients utilized for testing Surv-TransNet.

4.2 Standardized Image Pre-Processing Protocol

To get rid of scanner-induced artifacts, field inhomogeneities, and spatial misalignments across different imaging sites, we ran all 3D mpMRI volumes through a fully automated, standardized four-step pre-processing pipeline built with ITK and ANTs [5, 18, 33].
Step 1: N4ITK Bias Field Correction — We used Non-parametric Non-uniform intensity Normalization (N4ITK) to fix B1 magnetic field spatial intensity variations in all four MRI sequence modalities [18, 33].
Step 2: ITK Rigid Spatial Co-Registration — Native T1, T2, and T2-FLAIR volumes were all co-registered to the high-resolution post-contrast T1c reference space using ITK Mattes Mutual Information rigid registration (6 degrees of freedom) [5, 18].
Step 3: Isotropic Volumetric Resampling — Every 3D spatial volume was resampled to a uniform isotropic resolution of $1.0 \times 1.0 \times 1.0$ mm^3. We used 3D B-spline interpolation for the imaging channels and nearest-neighbor interpolation for the ground-truth masks [1, 5].
Step 4: Voxel Intensity Z-Score Normalization — We clipped extreme non-zero background voxel intensities at the 0.5th and 99.5th percentiles, then did sequence-specific Z-score normalization (zero mean, unit variance) [5, 18].
The algorithmic settings were tuned for each step as follows [5, 18, 33],
For N4ITK bias field correction (Step 1), convergence was defined as a maximum of 50 iterations over 4 multi-resolution fitting levels (100x50x30x20 iterations). We used a Gaussian smoothing kernel with FWHM of 0.15 mm and a spline distance of 200 mm [18, 33].
For rigid spatial co-registration (Step 2), post-contrast T1c was the fixed target reference. Native T1, T2, and T2-FLAIR acted as moving volumes [5, 18]. Optimization used a regular step gradient descent optimizer: minimum step size 0.001 mm, maximum step size 1.0 mm, and 32 histogram bins for the Mattes Mutual Information metric calculation [18, 33].
For isotropic resampling (Step 3), we applied 3D third-order B-spline interpolation to continuous intensity imaging channels to prevent ringing artifacts. For categorical ground-truth segmentation masks, we only used 3D nearest-neighbor interpolation to keep discrete label integrity intact [1, 5].
For Z-score intensity normalization (Step 4), intensity truncation at the 0.5th and 99.5th percentiles knocked out hyper-intense blood vessel artifacts and scanner dropouts before we computed the mean μ and standard deviation σ for intracranial brain tissue [5, 18].

4.3 Deep Learning Hardware & Software Infrastructure

Our work like model development, hyperparameter tuning, ablation experiments, and benchmarking were all done on a dedicated high-performance computing cluster. We have listed here the specifications we utilized:
- Compute: Four top-of-the-range NVIDIA A100 Tensor Core GPUs, each boasting 80GB HBM2e VRAM (so altogether, that makes 320GB VRAM in total) connected via NVLink for a quick 600 GB/s bidirectional data transfer rate.
- Host: our compute nodes had two high-end AMD EPYC 7763 CPUs with combined cores of 128 and threads of

256, respectively. We had a total of 512GB of system memory (DDR4) and 15TB of NVMe enterprise SSD scratch storage array.
- Software: our development stack consisted of Ubuntu operating system 22. 04 LTS, PyTorch structure at the 2. 2. 0 version, MONAI at 1. 3. 0, CUDA graphics computing environment at 12. 1, cuDNN libraries at 8. 9. 2 (deep learning primitives libraries for fast computation), and Python 3. 10 as the programming language.
We used all-four A100 devices together (multi-device training) to maximize GPU usage and reduce IO waits in distributed training. We also pre-loaded all the processed 3D isotropic mpMRI volumes into a RAM-disk (/dev/shm) to remove data pipeline bottlenecks. When it came to mixed precision (FP16) in PyTorch (torch. cuda. amp. autocast), this single step brought GPU memory usage by 44% reduction while making the forward-backward pass (gradient updates) up to 2. 3 times faster when compared to straight FP32. Dynamic loss scaling (`torch. cuda. amp. GradScaler`) safeguarded from loss of the smaller gradients during FP16 gradient backpropagation, mostly in the deep swin-transformers self-attention layers. Synchronizing the weight gradients over NVLink kept the scaling efficiency almost linear - 98. 2% on the four NVIDIA A100 GPUs.

4.4 Comprehensive Hyperparameter Search & Training Protocols

We trained the Surv-TransNet end-to-end with AMP FP16 to get the most out of the memory and improve the speed [5, 7]. Each model was trained for 300 complete epochs in each fold within a 5-fold cross-validation system, where the folds were stratified by volume and death/recovery status [1 4 5] [1, 4, 5]. Here's how we handled optimization:
- To optimize, we chose the AdamW, starting with a learning rate of 1 10 -4. Momentum parameters are set to 1 = 0. 9 and 2 = 0. 999. The weight decay is 1 10 -5 [5, 7].
- The learning rate was annealed per a cosine schedule over 20 epochs with a 20 epoch linear warm-up. It eventually waned to a minimum of 110 -7 [5, 7].
- (a) each batch, 128 128 128 voxels patches were randomly sampled from the tumor boundary with a tumor to non-tumor ratio of 1:1 [5, 18]. Batch size was 8 (2 samples per GPU) [5, 7].
- All data augmentation was done on the fly and included: Sample random 3D rotations (up to 15); scaling (0. 85 to 1. 15); elastic deformations (= 5, = 100); additive Gaussian noise (= 0. 1); Gaussian blur (between 0. 5 and 1. 5); and gamma correction (from 0. 7 to 1. 5) [5, 18].
Hyperparameters: for our hyperparameters, we used Bayesian optimization with Optuna, running 50 different search trials over the validation split [5, 7]. We optimized the loss function coefficients over the respective ranges: [0. 5, 2. 0], [0. 1, 1. 0], and [0. 1, 1. 5] [3 4 21] [3, 4, 21]. Ultimately, the optimal values of , , and were 1. 0, 0. 5, and 0. 8 respectively for regional segmentation, BPCL boundary loss, and Cox hazard loss [1 3 21] [1, 3, 21]. In the CAMF module: cross-attention dropout was set to 0.
1; we used 4 multi-head attention streams and projected clinical embeddings to 256 dimensions [4 29 32] [4, 29, 32]. The Deep Cox Hazard MLP head had three layers (256 128 1), GELU activations, and Dropout (0.
3)just enough to prevent overfitting on right-censored survival data [21, 26]. Finally, we used early stopping if the validation C-index didn't improve for 30 epochs in a row [4, 21].

4.5 Quantitative Evaluation Metrics

We used a series of solid statistical measures that evaluate model performance against the two target tasks [1 4 5, 21]: [1, 4, 5, 21]
1. It's Harrell's C - Index (C - Index) metric that evaluates the quality of the survival prediction ranking by the model [21, 26]. If C - index is 0. 5 means it is a random result, but if it is 1. 0 it is perfectly sorted, meaning that model perfectly identifies the longer living ones [21]. [...]
2. It is a Time-Dependent AUC12 and AUC24 that represent the areas under curve by ROC analysis for 12 and 24 months survival respectively [4, 28].
3. Survival Classification of 3-Class accuracy is a measurement that shows how well the model identifies the patients correctly and puts them in these groups of short (<10 mo), medium (18 mo) and long term (>18 mo) survivors [4, 28].
4. The tumor predicted tumor mask (Y_hat) is compared to the ground-truth mask (Y) through a metric called Dice Similarity Coefficient (DSC) which is reported as Whole-Tumor (WT), tumor core (TC) and enhancing tumor (ET) [1 5 6] [1, 5, 6].
5. It is a 95th Percentile of the maximum Hausdorff Distance (HD95) between tumor contours of prediction and ground-truth [1, 5, 18]
We used the standard formulas for all the metrics and this helped in making sure that results remain interpretable and valid not only from one model and one task to another but also across the entire set of tasks [1 4 5, 21] [1, 4, 5, 21].
Here's how one may get to the calculation of these measures:- Harrell's C-Index checks all patient pairs (i, j) where $T_i < T_j$ and $\delta_i = 1$:

C-index = (sum over all such pairs of $I(T_i < T_j) * I(g_i > g_j) * \delta_i$) / (sum over all such pairs of $I(T_i < T_j) * \delta_i$)

where $I(\cdot)$ is just an indicator function, and g_i, g_j are the model's predicted log-hazard scores [21].

- Dice Similarity Coefficient (DSC) for a given region k:
  DSC_k = (2 * |Y_k ∩ Y_hat_k|) / (|Y_k| + |Y_hat_k|)
- 95th Percentile Hausdorff Distance (HD95):
  HD95 = max_{95%}(max_{a ∈ ∂Y} min_{b ∈ ∂Y_hat} ||a-b||_2, max_{b ∈ ∂Y_hat} min_{a ∈ ∂Y} ||a-b||_2)

Finally, to see if the differences between models actually matter, we used paired two-tailed Student's t-tests and Wilcoxon signed-rank tests, setting our significance threshold at $p < 0.01$ [1, 4, 5].

5. Results & Comparative Analysis

5.1 Quantitative Survival Prediction Benchmarks

Table 1 presents the quantitative survival prediction benchmarks comparing Surv-TransNet against state-of-the-art survival paradigms across n = 2,140 multi-institutional patient cases [1, 3, 4, 21, 26, 28]. Surv-TransNet achieved a mean C-index of 0.812 for continuous overall survival prediction, significantly outperforming existing 3D U-Net + Linear Cox (0.642), DenseNet + DeepSurv (0.698), Swin UNETR + Radiomics (0.745), and TransUNet + Direct MLP (0.731) baselines [4, 7, 21, 28, 29, 30].

TABLE 1: Quantitative Overall Survival Prediction Benchmarks Across Cohorts (n = 2,140)

Model Architecture	Input Modalities	C-Index ↑	AUC12 ↑	AUC24 ↑	3-Class Acc (%) ↑
3D U-Net + Linear Cox	MRI	0.642	0.685	0.650	58.2%
DenseNet + DeepSurv	MRI + Age	0.698	0.732	0.704	64.1%
Swin UNETR + Radiomics	MRI + Age + KPS	0.745	0.781	0.752	67.4%
TransUNet + Direct MLP	MRI + Clinical	0.731	0.768	0.739	66.0%
Surv-TransNet (Ours)	MRI + Clinical + IDH/MGMT	0.812	0.864	0.831	72.8%

Surv-TransNet got time-dependent areas under the ROC curves as high as 0. 864 at a year and 0. 831 at a two-year mark. The classification accuracy of 3 classes came to 72. 8%. This is a great improvement from former methods. By incorporating multi-sequence 3D mpMRI data with genetic biomarkers through cross-attention, the result is a much larger predictive accuracy than just relying on traditional radiomics or single modality methods.

Examining various patient populations from different clinical studies, Surv-TransNet performed well, consistently. For BraTS 2023 Glioma Benchmark split data consisting of n = 1,250 cases, the C-index of model was 0. 805. For BraTS 2021 Survival Challenge benchmark having n = 546, the model's scoring was a C-index of 0. 818, surpassing the top ensemble model of the challenge by 0. 056. The model shone the most at the CHIC-344 cohort (n = 344), where complete molecular profiles (IDH1/2 and MGMT) were provided. Here, Surv-TransNet attained its best: a C-index of 0. 826, and a 12-month area under the curve (AUC) of 0. 881.

When predicted hazard scores were categorized into high, medium, and low-risk groups, survival curves drawn with Kaplan-Meier were much different and this was confirmed by a log-rank test with $p < 0.0001$. On average, survivors with low-risk feature lasted 624 days, medium-risk ones lasted 380 days, and high-risk ones lasted just 182 days, which clearly supports the results of hazard estimates being the true picture survival outcome of model Surv-TransNet.

5.2 Multi-Class 3D Tumor Segmentation Performance

Table 2. 3D tumor segmentation results for the semantic subregions: Whole Tumor (WT), Tumor Core (TC) and Enhancing Tumor (ET) [1, 5, 6]. Surv-TransNet achieved the state-of-the-art segmentation accuracy with Dice Similarity Coefficients of 93.85% for WT, 91.20% for TC, and 89.40% for ET, with a low 95th percentile Hausdorff Distance (HD95) of 2.38 mm [1, 5, 7, 18, 31].

TABLE 2: Multi-Class 3D Tumor Segmentation Accuracy Across Subregions

Model Architecture	Whole Tumor (WT) DSC (%)	Tumor Core (TC) DSC (%)	Enhancing Tumor (ET) DSC (%)	HD95 (mm) ↓
3D U-Net	88.45	84.12	82.30	6.42

SegResNet	91.20	87.60	85.45	4.95
Swin UNETR	92.85	89.40	87.10	3.82
Surv-TransNet (Ours)	93.85	91.20	89.40	2.38

The Spatial Edge Attention (SEA) modules and the Boundary-Preserving Contrastive Loss (BPCL) techniques have been quite effective in preventing the erosion of boundaries around the area of the invasive edema. What this results in is a better sharpness of the edges and much higher HD95 scores when compared to the standard 3D U-Net (6. 42 mm) or Swin UNetR (3. 82 mm) backbones [3 5 7 10 31] [3, 5, 7, 10, 31].

A closer look at the subregion segmentation results on the most challenging pathological instances reveals that Surv-TransNet is genuinely spatially stable [1 2 5] [1, 2, 5]. For the infiltrating and non-contrast-enhancing diffuse tumors example, typical deep CNNs have difficulties recognizing and defining the boundaries while Surv-TransNet has still been managing the Whole Tumor (WT) DSC at 92. 4% [5 6 10] [5, 6, 10]. For the most complex cases, like the very necrotic, heterogeneous glioblastomas, the Netrotic Core (NET) DSC remained at 88. 6% [1, 5, 14]. The ring of contrast-enhancing post-contrast tumor (ET) was scored 89. 4 [1 5 14].

When we actually visualize the predicted 3D-segmentation masks overlapped with post-contrast T1c and T2-FLAIR imaging, it turns out: SEA modules really do get rid of the islands' artifacts that are disturbing, and they minimize the spreading of the perimeter which is commonly a flaw in standard Swin Transformer self-attention blocks [7 10 31] [7, 10, 31]. The changes here extend far more than superficial improvements alone. Because of the study, the 95th percentile of the Hausdorff Distance (HD95) dropped remarkably for subregions and boundary errors were reduced by 1. 44 mm (from a baseline Swin UNETR) and 4. 04 mm relative to an original 3D U-Net [5 7 18]) [5, 7, 18].

5.3 Systematic Component Ablation Experiments

We perform a systematic ablation study to analyze the individual performance contribution of key architecture components, as shown in Table 3 [3, 4, 7, 10, 32].

TABLE 3: Systematic Component Ablation Breakdown

Configuration	C-Index ↑	WT DSC (%) ↑	AUC12 ↑
Baseline (Swin Encoder + MLP Loss)	0.712	91.10	0.742
+ Spatial Edge Attention (SEA)	0.735	92.80	0.771
+ BPCL Loss (L_BPCL)	0.758	93.85	0.795
+ CAMF Module (Full Multimodal Fusion)	0.812	93.78	0.864

Ablation results show the clarity: the Swin-Transformer backbone is a strong initial baseline (C-index 0. 712, WT DSC 91. 10%). And when Spatial Edge Attention is incorporated, the metrics increase a 0. 023 gain in C-index. With the addition of BPCL loss, the same 0. 023 C-index growth is achieved with a 1. 05% gain in WT DSC. These improvements are not unsound increments, but immediate improvements in spatial acuity.

Then there is the huge jump. As soon as you go to the CAMF multimodal cross-attention module, you get the biggest single leap in prognostic poweryour C-index jumps another 0. 054, and your 12-month AUC jumps another 0. 069Sounds pretty promising. You may have found the system that truly combines radiogenomic features effectively.

Breaking this into steps reveals the story even more plainly. The baseline 3D Swin-UNet with a mere concatenation of vectors for survival prediction achieves a C-index of 0. 712 and WT DSC of 91. 10%.

Introducing a SEA to the skip connections itself increases WT DSC by 1. 7% (92. 80%) and the C-index of the survival task by 0. 023 to 0. 735. This means explicitly focusing the model on spatial gradientsthe modeling of more accurate tumor boundarieshas a positive effect on survival risk prediction. Then, when the BPCL plays a role again, we can see that the WT DSC gets back to 93. 85% and the C-index gets back to 0. 758.

Why? Because BPCL commands the encoder to learn more on the tissue parcels surrounding tumor edges to maintain boundary persistence. Finally, using the CAMF cross-attention module instead of the simple vector concatenation yields a considerable improvement in the clinically relevant prognostics power of the model. The C-index jumps to 0. 812 and the 12-month AUC climbs to 0.

864, with excellent segmentation accuracy (WT DSC: 93. 78%). In other words, in prediction of survival for high grade gliomas, meaning a good spatial delineation combined with an efficient method to fuse dynamically the molecular genetic data is required becausethe two components go hand in hand. Improving one about the other isn't effective.

6. Discussion & Clinical Implications

6.1 Methodological Advances & Elimination of Decoupling Bias

These trials have convincingly shown that combining spatial boundary extraction of the tumor shape in 3D with clinical-genomic cross-attention really is a good solution to the Decoupling problem that radiomics workflows typically have. Two-stage methods from the past just accumulate errors, if segmentation is incorrect even minutely, mistakes carry over to and mess up all the downstream radiomic features. Then again, Surv-TransNet trains both segmentation decoders and survival hazard heads simultaneously in one multi-task setting. As a result, it forces the encoder to retain and focus on spatial details from the microenvironment, for example, fine gradients of the tumor outline or abnormalness in small blood vessels. Such features in fact have been very important in predicting the outcomes of patients.

A major strength at the conceptual level of Surv-TransNet is its ability to minimize loss of information at the edges where the cancer cells have started invading. Generally, radiomic pipelines discard the non-tumor regions after the segmentation network makes a binary mask. For instance, for high-grade gliomas, research shows that the tumor cells actually extend a lot beyond the edge or can even be present among "unaffected" tissue and vasogenic edema up to 20 mm away from the tumor. WithSurv-TransNet's end-to-end gradient pathway, survival modeling guides the spatial features throughout the 3D scan, and because of this, information in the peritumoral areas with their tumor infiltration and the infiltrative fronts are not lost. The shared Swin-Transformer bottleneck so gradually detects tumor infiltration signs and vascular variation that are non-local, which traditional segmented radiomics completely fail to capture.

6.2 Synergistic Radiogenomic Fusion via Cross-Attention

The CAMF module also effectively incorporates high-dimensional imaging data with low-dimensional clinical/genomic data, like age KPS IDH1/2, and MGMT. The attention maps show that spatial imaging data gives the initial anatomical structure, but the molecular markers of course, most strongly influence the shape of the risk curves, which is likely reflected in the slant of the predicted risk curves.

This type of cross-talk between imaging and genomics just makes clinical sense. Imaging shows you the macro snap-shotmass effect edema what is going on anatomically in the brain, while the molecular perspective overlaid on that graph brings all information about cells' potential aggressiveness, drug and radiation response so.

The actual attention weights CAMF produces aren't just a confusion-based artifact. They can also shed light biologically on the images, are reliant on the information broken down by patient subgroup. In IDH1/2-wildtype glioblastomas (more malignant, hypervascular tumors), the attention is in the right placeon the necrotic core and the enhancing rim on the post-contrast T1c.

This is what we want, these are both features linked to increased survivability for these tumors and believeable breeding ground for tumor necrosis. Step away from IDH1/2-wildtype glioblasomas and their attention moves to the T2-FLAIR edema, a manifestation of the tumors' more diffuse, less aggressive growth. With MGMT promoter methylation, CAMF draws on clinical-genomics features more heavily, mainly when it is trying to surmise who will live over a year.

That is consistent with the ameliorated response to temozolomide in this group. So, CAMF isn't just doing numbers or performing vector matching but is actually capturing the thought process of neuro-oncologists when about tumor biology. The module does not just function but is thinking like clinicians.

6.3 Clinical Translation & Decision Support in Neuro-Oncology

In real-life scenarios, Surv-TransNet's pre-operative survival hazard curves give neuro-oncology teams a solid, data-based method to justify their decisions. With the help of very accurate and continuous survival hazard scores, the surgical teams can adjust their plans very precisely. They decide how much of the tumor to remove, carefully considering the pros and cons of aggressive surgery versus preserving brain function, and identify high-risk patients who should be offered early-stage trials or targeted therapies first.

Fine-tuning the risk models to be part of hospital operations means that the tools will be very well integrated with existing hospital systems like PACS and EHR platforms. Surv-TransNet runs as a containerized microservice, it can pull DICOM mpMRI directly from PACS, perform all processing and inference, and produce standardized DICOM-SR reports without the user knowing.

In the surgery room, Surv-TransNet's 3D contours can be displayed right on neurosurgery tools like BrainLab or StealthStation thereby assisting fluorescence-based resections (like 5-ALA) or tracking interoperative ultrasound. Like that, the model-generated continuous survival risk curves and 12- or 24-month survival probabilities are also accessible directly in the EHR dashboards. In this way, the tumor board can get together and, with just one quick look at all the data, decide how best to proceed - clear and interpretable radiogenomic profiles that will support them in determining the treatment sequence that will follow.

7. Conclusion

Surv-TransNet is our end-to-end deep learning model for simultaneous 3D tumor segmentation and survival

prediction of high-grade gliomas. We assembled a number of important components - Swin-Transformer encoders, Spatial Edge Attention, CAMF for multimodal fusion and a Deep Cox Hazard neural head - so that the system integrates well and does not suffer from the usual compounding errors. Surv-TransNet was evaluated on 2,140 scans from different centers and achieved new state-of-the-art results for survival prediction (C-index 0.812) and 3D tumor segmentation accuracy (Whole Tumor DSC 93.85%). Seems like a good step forward in personalized neuro-oncology care

References

1. Bakas, S., et al. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data, 4, 170117.
2. Louis, D. N., et al. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology, 23(8), 1231–1251.
3. Menze, B. H., et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging, 34(10), 1993–2024.
4. Katzman, J. L., et al. (2018). DeepSurv: personalized treatment recommendation system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology, 18(1), 24.
5. Isensee, F., et al. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211.
6. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI, 234–241.
7. Hatamizadeh, A., et al. (2022). Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. IEEE Access, 10, 27299–27309.
8. Stupp, R., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine, 352(10), 987–996.
9. Weller, M., et al. (2021). EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nature Reviews Clinical Oncology, 18(3), 170–186.
10. Tang, Y., et al. (2022). Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. IEEE CVPR, 20730–20740.
11. Hegi, M. E., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. New England Journal of Medicine, 352(10), 997–1003.
12. Yan, H., et al. (2009). IDH1 and IDH2 mutations in gliomas. New England Journal of Medicine, 360(8), 765–773.
13. [13] Marcus, D. S., et al. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience, 19(9), 1498–1507.
14. LaBella, D., et al. (2024). The ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2023: Intracranial Meningioma. Radiology: Artificial Intelligence, 6(2), e230170.
15. Kamnitsas, K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis, 36, 61–78.
16. Myronenko, A. (2018). 3D MRI brain tumor segmentation using autoencoder regularization. Brainlesion Workshop, Springer, 311–320.
17. Chen, J., et al. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306.
18. Tustison, N. J., et al. (2010). N4ITK: Nick's N3 ITK implementation for MRI bias field correction. IEEE Transactions on Medical Imaging, 29(6), 1310–1320.
19. Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images Are More than Pictures, They Are Data. Radiology, 278(2), 563–577.
20. Aerts, H. J., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications, 5, 4006.
21. Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202.
22. Zhou, Z., et al. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. DLMIA Workshop, Springer, 3–11.
23. Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 3D Vision (3DV), 565–571.
24. Zwanenburg, A., et al. (2020). The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image Analysis. Radiology, 295(2), 328–338.
25. Lambin, P., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology, 14(12), 749–762.
26. Ching, T., et al. (2018). Cox-nnet: An artificial neural network method for prognosis prediction for cancer patients. PLOS Computational Biology, 14(4), e1006076.
27. Kickingereder, P., et al. (2016). Radiomic profiling of glioblastoma: identifying an imaging signature

associated with survival. Neuro-Oncology, 18(9), 1284–1292.
28. Feng, X., Tustison, N. J., & Meyer, C. H. (2020). Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using random forests. Brainlesion Workshop, Springer, 279–288.
29. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008.
30. Tang, Y., et al. (2021). High-resolution 3D medical image segmentation with transformers. arXiv preprint arXiv:2103.01831.
31. Liu, Z., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV, 10012–10022.
32. Huang, S. C., et al. (2021). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digital Medicine, 4(1), 136.
33. Avants, B. B., et al. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage, 54(3), 2033–2044.