

ProtoSurv-X: Explainable Brain Tumor Survival Prediction

Neeshu Kumari¹ , Saurabh Shah¹ , Chintan Thacker² 

¹ Computer Science and Engineering, Parul University, Gujarat, India

² Computer Science and Engineering, Kachchh University, Gujarat, India

Corresponding author: Neeshu Kumari (e-mail: 212300470010@paruluniversity.ac.in), **Author(s) Email:** Saurabh Shah (e-mail: director.tcd@paruluniversity.ac.in), **Author(s) Email:** Chintan Thacker (e-mail: chintanthacker450@gmail.com).

Abstract Accurate survival prediction for patients with high-grade glioma is important for prognostic assessment and personalized treatment planning; however, substantial intratumoral heterogeneity, complex multimodal MRI patterns, and limited interpretability challenge existing deep learning approaches. This study aims to develop ProtoSurv-X, an explainable and uncertainty-aware framework for MRI-based glioma survival prediction that integrates probabilistic tumor habitat modeling, prototype-guided learning, evidential prediction, and evidence-grounded clinical explanations. A unified cohort was constructed from the BraTS 2019 and BraTS 2020 datasets by removing duplicate subjects, resulting in 369 unique subjects, including 118 gross total resection patients with complete survival annotations. The proposed framework uses diffusion-enhanced SwinUNETR for tumor segmentation, probabilistic habitat construction, and fusion of radiomic, deep imaging, habitat, and age-related features. Prototype learning and survival-aware contrastive learning generate prognostic representations, while Evidential Deep Learning estimates risk and predictive uncertainty alongside continuous survival regression. For explanation generation, seven open-source large language models were benchmarked using structured model evidence, and a Meta-Llama-3.1-8B-Instruct model was fine-tuned using QLoRA. In five-fold cross-validation on the unified survival cohort, ProtoSurv-X achieved a mean absolute error of 145.2 ± 6.5 days, an RMSE of 189.4 ± 8.3 days, a C-index of 0.745 ± 0.008 , and an integrated Brier score of 0.124. The segmentation module achieved Dice scores of 91.24%, 87.38%, and 81.76% for whole tumor, tumor core, and enhancing tumor, respectively, on BraTS 2019. The fine-tuned explanation model obtained a BERTScore of 0.915 and a hallucination rate of 2.6%. These findings indicate that ProtoSurv-X offers a unified computational framework for accurate, uncertainty-aware, and evidence-grounded glioma prognostic modeling, while further clinical and multicenter validation remains necessary.

Keywords: Glioma Survival Prediction; Diffusion-Enhanced SwinUNETR; Tumor Habitat Modeling; Prototype Learning; Evidential Deep Learning; Clinical Reasoning

1. Introduction

Gliomas are highly aggressive brain tumors characterized by profound intra-tumoral heterogeneity, making accurate pre-operative survival prediction essential for personalized treatment planning and risk stratification. While multi-parametric magnetic resonance imaging (MRI) is the primary diagnostic modality, predicting survival remains challenging because prognostic information is distributed across complex, localized tumor sub-regions. Although deep learning architectures like SwinUNETR achieve state-of-the-art tumor segmentation on the BraTS benchmark, conventional pipelines generate hard, deterministic boundary masks that discard critical spatial uncertainty. Transitioning from deterministic masks to probabilistic tumor habitat modeling directly

addresses this gap by preserving anatomical ambiguity and infiltrative uncertainty at tumor boundaries, providing downstream prognostic models with a higher-fidelity representation of intra-tumoral aggressiveness. MRI-based survival prediction has recently received considerable attention. Early approaches relied on handcrafted radiomic features and conventional machine learning, whereas recent methods employ deep learning, graph neural networks (GNNs), and transformer architectures to learn prognostic representations directly from imaging data. For example, MTC-Net combines CNNs and transformers for joint segmentation and survival prediction, while cooperative GNNs model structural relationships across MRI modalities to improve prognostic performance [1,2]. Likewise, multimodal prototype learning frameworks, such as Multimodal

Prototyping for Cancer Survival Prediction (MMP), demonstrate the effectiveness of prototype-based representations for interpretable survival modeling across heterogeneous biomedical data [3]. However, these deep learning architectures predominantly rely on global image representations or whole-tumor feature pooling, which fundamentally fails to capture intra-tumoral complexity. By averaging all regions together, existing models fail to capture the habitat-specific patterns needed for accurate risk stratification. Moreover, prototype-based methods such as MMP were mainly developed for 2D pathology images and do not effectively model the complex multi-modal interactions present in 3D brain MRI. Reliable oncology prognostication requires both accurate uncertainty estimation and clear interpretability. Although evidential models such as DPSurv [4] estimate uncertainty using disease prototypes, they were developed mainly for 2D pathology images and are not designed for high-dimensional 3D multi-parametric MRI. Saliency maps highlight image regions but do not explain the clinical basis of predictions. Similarly, although Large Language Models (LLMs) can generate fluent reports, they often produce explanations that are not grounded in quantitative model evidence, increasing the risk of hallucinations and limiting their reliability for survival prediction.

To address these limitations, we propose ProtoSurv-X, a unified prognostic framework that integrates probabilistic tumor habitat modeling, prototype-guided survival learning, and evidence-grounded natural language explanation. This study aims to develop an interpretable MRI-based glioma survival prediction framework that integrates probabilistic tumor habitat modeling, prototype-guided survival learning, uncertainty estimation, and evidence-grounded LLM explanations. The main contributions of this work are summarized as follows:

1. We introduce a habitat-aware prototype learning strategy that discovers clinically meaningful prognostic archetypes from fused radiomic, deep imaging, habitat, and clinical representations.
2. We jointly optimize survival-aware contrastive learning and Evidential Deep Learning to improve continuous survival prediction while providing calibrated predictive uncertainty and improved prognostic interpretability.
3. We implement an evidence-grounded LLM explanation generator that combines zero-shot model selection with QLoRA-based instruction fine-tuning, translating deterministic survival metrics into reliable, hallucination-resistant clinical reports

II. Related Work

A. Brain Tumor Segmentation

Automated brain tumor segmentation is a fundamental task in neuro-oncology, providing accurate delineation of tumor regions for diagnosis, treatment planning, and quantitative analysis. Early deep learning methods, including U-Net proposed by Ronneberger et al [5], V-Net proposed by Milletari et al [6], and nnU-Net developed by Isensee et al. [7], established strong baselines; SwinUNETR, proposed by Hatamizadeh et al. [8], significantly improved multi-scale feature learning and segmentation performance. Representative methods include M3AE proposed by Zhang et al. [9], SMML proposed by Liu et al [10], and the prompt-guided framework of Zhao et al. [11], which combine modality reconstruction, self-supervised learning, and semantic guidance to maintain segmentation accuracy under incomplete imaging conditions. Recent studies have further explored advanced representation learning strategies, including modality decomposition and graph-based reasoning. Wang et al.'s KMD [12] disentangles shared and modality-specific representations, while Tao et al. [13] employ graph neural networks to model relationships among multiple MRI modalities.

B. MRI-Based Survival Prediction

Recent advances have shifted toward deep learning-based survival prediction frameworks that learn prognostic representations directly from multiparametric MRI data. Mazumdar and Mukhopadhyay proposed [14] the Multi-Scale Transformer CNN Network (MTC-Net), which integrates convolutional feature extraction. Similarly, Datta et al [15] introduced a multimodal exponential moving average guided framework that jointly integrates MRI-based tumor segmentation and survival prediction through knowledge distillation, thereby mitigating the limitations associated with disjointed radiomic pipelines. Cao et al. [1] proposed a Cooperative Multiplex Graph Neural Network (GNN) that models intra- and inter-modal relationships from multi-modal MRI to improve glioma survival prediction. Song et al. [3] proposed Multimodal Prototyping for Cancer Survival Prediction (MMP), which summarizes large scale imaging and omics representations. Qu et al. [16] introduced MDCS-MoAME, leveraging multimodal pathology and genomic information through Mamba-based representation learning.

C. Prototype Representation Learning

Prototype learning represents patients using representative disease archetypes rather than opaque latent features. Early studies, such as Shao et al.'s [17] PANTHER, demonstrated that morphological prototypes can effectively summarize heterogeneous tissue patterns, improving classification, survival prediction, and interpretability in computational pathology. Wu et al. [18] proposed a Pathology-Aware

Mixture-of-Experts (PAMoE) framework that learns tissue specific experts capable of specializing in distinct intratumoral structures. Song et al. [3] extended morphological prototyping to the multimodal domain(MMP), combining WSI prototypes with biological pathway prototypes derived from transcriptomic data to reduce computational complexity. Li et al. [4] introduced the DPsurv framework, a dual-prototype evidential fusion architecture for WSI survival prediction, employing prototype assignment mechanisms to identify representative tissue patterns.

D. LLM-Based Clinical Explanation

Large Language Models (LLMs) have recently emerged as powerful tools for clinical decision support, medical question answering, report generation, and multimodal reasoning. Among these developments, Li et al. [19] developed LLaVA-Med, which extended vision-language instruction tuning to the biomedical domain by enabling conversational reasoning over medical images and visual question answering tasks. Similarly, Wu et al. [20] introduce RadFM, a large-scale radiology foundation model capable of jointly processing 2D and 3D medical imaging data for diagnosis and report generation. Kandala and GVS [21] proposed an interpretable diagnostic pipeline combining concept bottleneck learning, SHAP-based explanations, and LLM-driven clinical reasoning generation. Li et al. [22] demonstrated that LLM-based evaluators frequently achieve moderate-to-strong agreement with human experts across clinical decision support, medical question answering, and clinical text generation tasks, providing a scalable alternative for evaluating complex clinical narratives. To address these limitations, we propose a unified MRI-based glioma prognosis framework integrating tumor habitat modeling, prototype-guided survival prediction, evidential uncertainty estimation, and evidence-grounded LLM-based clinical explanation generation.

III. Methods

ProtoSurv-X is a five-stage framework for interpretable MRI-based glioma survival prediction, comprising MRI preprocessing, Diffusion-Enhanced SwinUNETR tumor segmentation, habitat-aware feature extraction, prototype-guided evidential survival prediction with contrastive learning, and LLM-based, evidence-grounded clinical explanation generation, as shown in Fig. 1.

A. Dataset and Preprocessing

The survival cohort was constructed following the official BraTS survival prediction protocol. Only patients with Gross Total Resection (GTR), complete overall

survival (OS), and age information were included. Duplicate subjects between the cumulative BraTS 2019 [2] and BraTS 2020 [23] datasets were removed using patient identifiers, resulting in a unified cohort of **369 unique imaging subjects**, of which **118 GTR subjects** were eligible for survival prediction. The final dataset was split into patient-level training, validation, and test sets to prevent data leakage and ensure unbiased evaluation.

B. Tumor Segmentation via Diffusion-Enhanced SwinUNETR

Given a multi-parametric MRI volume $X = \{X_{T1}, X_{T1ce}, X_{T2}, X_{FLAIR}\}$, we integrate a lightweight diffusion-based feature enhancement module directly before the SwinUNETR patch embedding encoder to refine structural boundaries and suppress acquisition noise prior to representation learning. In the forward

diffusion process, Gaussian noise is iteratively added to the input imaging features over $T = 1000$ timesteps using a cosine noise schedule [35] $\beta_T \in [10^{-4}, 0.02]$ as illustrate in Eq. (1) [36]

$$q(x_t|x_0) = N(x_t; \sqrt{\alpha_t}x_0(1 - \alpha_t)I) \quad (1)$$

where x_0 is the original MRI feature volume, x_t is the noisy feature at timestep t , α_t is the cumulative noise-retention coefficient, ϵ is Gaussian noise, t is the diffusion timestep, and $T = 1000$ is the total number of timesteps. Where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. A denoising network c_0 is trained to predict the injected noise by minimizing the standard diffusion objective, which is jointly optimized with the segmentation loss, as computed in Eq. (2) [35].

$$L_{diff} = E_{x_0, \epsilon, t} [\|c - c_0(x_t, t)\|^2] \quad (2)$$

During inference, a 20-step DDIM sampling generates an enhanced volume. The hierarchical SwinUNETR processes $\hat{x}$ to output voxel-wise soft probability maps for the enhancing tumor (ET), necrotic core (NCR), and edema (ED), denoted

$\hat{P} = \{\hat{P}_{ET}, \hat{P}_{NCR}, \hat{P}_{ED}\}$ where $\hat{p}_k(v) \in [0, 1]$, for $k \in \{ET, NCR, ED\}$, represents the predicted probability that voxel v belongs to the corresponding tumor sub-region.. To propagate segmentation uncertainty into prognostic modeling, clinically relevant regions are probabilistically aggregated and bounded to $[0, 1]$ for the whole tumor (WT), tumor core(TC), and Enhancing Tumor (ET) in Eq. (3) [2] [23] where $\hat{P}_{WT}$ and $\hat{P}_{TC}$ denote the probabilistic whole-tumor and tumor-core maps at voxel vv , respectively.

$$\hat{P}_{WT} = \min(1, \hat{P}_{ET} + \hat{P}_{NCR} + \hat{P}_{ED})$$

$$\hat{P}_{TC} = \min(1, \hat{P}_{ET} + \hat{P}_{NCR}) \quad (3)$$

C. Tumor Habitat Construction and Multi-Stream Feature Extraction

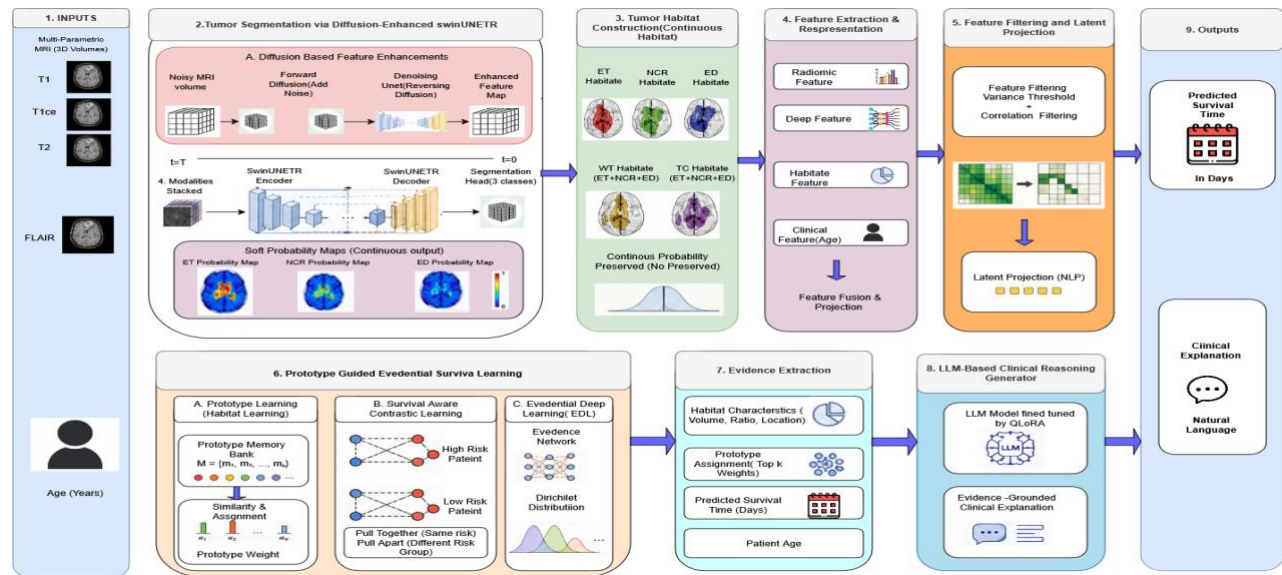


Fig. 1. Overview of the proposed ProtoSurv-X framework. Multi-parametric MRI and patient age are processed through diffusion-enhanced tumor segmentation, probabilistic tumor habitat construction, multimodal feature extraction, prototype-guided evidential survival prediction, and evidence-grounded LLM-based clinical explanation generation.

To capture intratumoral heterogeneity, the soft segmentation maps are utilized to define distinct morphological “habitats”, including the active enhancing region, the necrotic core, the surrounding edema, and the biologically critical tumor boundary margin. Enhancing, necrotic, and edema habitats were derived from probabilistic segmentation maps, while the boundary-margin habitat was extracted from tumor probability edges. Probabilistic aggregation preserved overlapping characteristics. For each patient, features from four domains were fused into a unified multi-stream vector Radiomics Features (f_{rad}): Extracted using PyRadiomics, capturing shape, first-order statistics, and texture matrices from the regional probability maps. Deep ViT Features (f_{deep}): High-dimensional spatial embeddings extracted from the SwinUNETR bottleneck. Habitat Features (f_{hab}): Volumetric and intensity interaction metrics between the defined tumor habitats. Clinical Features (f_{clin}): Patient age, serving as a foundational demo-graphic prognosticator. The initial patient representation is obtained via concatenation shown in Eq.(4) [3].

$$F_{concat} = f_{rad} \oplus f_{deep} \oplus f_{hab} \oplus f_{clin} \quad (4)$$

where F_{concat} denotes the concatenated feature vector and $\oplus$ denotes concatenation.

D. Multimodal Feature Filtering and Latent Projection

Because F_{concat} contains high dimensional SwinUNETR embeddings alongside exhaustive multi-regional radiomic texture matrices, the dimensionality of the concatenated feature vector can substantially

exceed the number of available survival cases ($D_{in} \gg N$). To reduce redundancy and overfitting, filtered radiomic features and multimodal representations are projected into a compact prognostic latent space through an MLP bottleneck with BN, Dropout, and LeakyReLU (denotes Leaky Rectified Linear Unit activation function) Eq.(5) [3]

$$F = Dropout \left(LeakyReLU \left(BN(W_p F_{concat} + b_p) \right) \right) \quad (5)$$

Where BN represents Batch Normalization, $W_p \in \mathbb{R}^{d \times D_{in}}$ and $b_p \in \mathbb{R}^d$ are learnable projection parameters, and D_{in} represents the dimension of the target latent space. This reduction yields the final, dense patient representation $F \in \mathbb{R}^d$. The parameters of the projection block (W_p, b_p) are optimized jointly with the prototype learning and survival prediction objectives, allowing the latent representation F to dynamically adapt to prognostic information during training.

E. Evidential Survival Prediction with Prototype Learning

The proposed framework performs survival prediction using a prototype-guided evidential learning strategy. The objective is to learn representative prognostic patterns from the training cohort while simultaneously quantifying predictive uncertainty. To achieve this, Prototype Learning is first employed to discover latent survival-related disease archetypes, after which Evidential Deep Learning is utilized to estimate survival risk and associated uncertainty.

1. Prototype Learning Module

Given the dense patient representation $F \in \mathbb{R}^d$ obtained from the latent projection block, the proposed framework learns a prototype memory bank consisting of K latent

prototypes, as computed in Eq.(6) [17]

$$M = \{m_1, m_2, m_3, \dots, m_K\}, \quad m_k \in \mathbb{R}^d \quad (6)$$

where K denotes the number of prototypes, with each prototype capturing a characteristic morphological and prognostic pattern observed in the training population. To determine the relationship between a patient and the learned prototypes, cosine similarity is computed in Eq. (7) [4] between the patient representation and each prototype:

$$s_k = \frac{F^T m_k}{\|F\| \|m_k\| + \epsilon}, \quad k = 1, 2, 3, \dots, K \quad (7)$$

where a small constant $\epsilon = 10^{-8}$ is added to the denominator for numerical stability. The similarity scores are normalized using a softmax operation, as illustrated in Eq. (8) [4]:

$$\alpha_k = \frac{\exp(s_k)}{\sum_{j=1}^K \exp(s_j)} \quad (8)$$

where α_k represents the contribution of prototype m_k to the current patient. A prototype-guided representation in Eq. (9) [4] is then constructed as a weighted aggregation of all prototypes:

$$Z = \sum_{k=1}^K \alpha_k m_k \quad (9)$$

Finally, the original patient representation and the prototype-guided representation are concatenated. To maintain dimensional consistency for downstream modules, this concatenated representation is passed through a linear projection layer to obtain the final prognostic embedding H that denotes the final prognostic embedding in Eq.(10) [4].

$$H = W_h[F|Z] + b_h, \quad H \in \mathbb{R}^d \quad (10)$$

where $[F|Z]$ denotes feature concatenation, and $W_h \in \mathbb{R}^{d \times 2d}$ and $b_h \in \mathbb{R}^d$ are learnable projection parameters. This formulation encourages patients with similar tumor morphology, habitat composition, and clinical characteristics to be mapped toward common prognostic prototypes, thereby improving interpretability and facilitating clinically meaningful explanation generation.

2. Prototype Memory Optimization

To ensure that the memory bank M learns biologically and clinically meaningful archetypes rather than collapsing into arbitrary latent vectors, the prototypes are jointly optimized with the network parameters. The prototype optimization loss L_{proto} is explicitly formulated as the sum of a clustering loss and a diversity loss, as derived in Eq. (11) [4]

$$L_{proto} = L_{cluster} + \gamma L_{div}, \quad (11)$$

where γ is a balancing hyperparameter. The clustering loss is calculated as in Eq. (12) [4]

$$L_{cluster} = \frac{1}{N} \sum_{i=1}^N \|F_i - Z_{i2}\|^2 \quad (12)$$

Minimizes the mean squared error between the original patient feature F_i and its prototype-reconstructed representation Z_{i2} , and N is the number of patients, forcing the prototypes to faithfully anchor the actual data distribution. To prevent mode collapse, the diversity loss minimizes the cosine similarity between distinct prototypes, ensuring they capture diverse and non-redundant tumor morphologies, as derived in Eq.(13) [4] where L_{div} denotes the diversity loss.

$$L_{div} = \sum_{j=1}^K \sum_{k \neq j}^K \frac{m_j m_k}{\|m_j\| \|m_k\| + \epsilon} \quad (13)$$

3. Survival-Aware Contrastive Learning

To enhance prognostic representation learning, a survival-aware contrastive learning strategy is incorporated to encourage patients with similar survival outcomes to occupy neighboring regions in the latent space. Let $H_i, H_j \in \mathbb{R}^d$ denote the prototype-guided embeddings of patients i and j . To elegantly capture the continuous nature of survival times rather than relying on hard positive and negative thresholds, we employ a continuous survival weighting scheme. $w_{ij} = \exp\left(-\frac{|y_i - y_j|}{\sigma}\right)$, $w_{ij} \in [0, 1]$ [31].

where y_i and y_j denote the observed survival times in days, and σ is a scaling parameter controlling temporal sensitivity. The survival-aware contrastive loss is then formulated as in Eq.(14) [31] [34]

$$L_{SCL} = - \sum_{i=1}^N \log \frac{\sum_{j \neq i} w_{ij} \exp\left(\frac{\text{sim}(H_i, H_j)}{\tau}\right)}{\sum_{k \neq i} \exp\left(\frac{\text{sim}(H_i, H_k)}{\tau}\right)}, \quad (14)$$

where $\text{sim}(H_i, H_j) = \frac{H_i^T H_j}{\|H_i\| \|H_j\| + \epsilon}$ computes the cosine similarity, and τ is the temperature parameter controlling the latent distribution to encourage survival-consistent clustering.

4. Evidential Uncertainty Prediction

Instead of generating deterministic predictions, the Evidential Deep Learning (EDL) module learns a distribution of evidence for each risk category, enabling simultaneous risk prediction and uncertainty estimation without requiring Monte Carlo sampling or ensemble models. Given the prototype-guided representation H , C denotes the evidence assigned to risk category, an evidential prediction head f_ϕ produces a non-negative evidence vector $e = f_\phi(H) \geq 0$ [20]. where $f_\phi(H)$ denotes the evidential prediction network parameterized by ϕ . The corresponding Dirichlet parameters are obtained as $\alpha_c = e_c + 1$ [41] for each risk category $c \in \{1, \dots, C\}$. The total evidence is computed as $S = \sum_{c=1}^C \alpha_c$ [22] yielding the predicted risk probability for class C is $p_c = \frac{\alpha_c}{S}$. To obtain a

continuous and monotonic prognostic assessment, the expected risk score R , S denotes the total accumulated evidence and predictive uncertainty U , are computed as in Eq. (15) [41].

$$R = \sum_{c=1}^C cp_c, \quad U = \frac{C}{S} \quad (15)$$

For evidential learning, patients were categorized into three clinically meaningful risk groups: high-risk (overall survival <300 days), medium-risk (300–450 days), and low-risk (>450 days), following the BraTS survival protocol. The risk-category boundaries were determined using overall survival together with the available BraTS clinical information (patient age) to provide clinically relevant prognostic stratification. The evidential prediction head performs risk classification to estimate predictive uncertainty, while a separate regression head independently predicts continuous overall survival time in days. The evidential branch predicts the corresponding risk category and its associated uncertainty via the Dirichlet distribution, serving as a confidence measure where larger values of U indicate higher predictive uncertainty. In addition to risk stratification, a separate regression head independently predicts the continuous overall survival time $\hat{y} = g_\psi(H)$ where $\hat{y}$ denotes the estimated overall survival in days. Consequently, the proposed framework simultaneously produces $\{\hat{y}, R, U\}$.

F. LLM-Based Clinical Explanation

To generate evidence-grounded survival explanations, we propose a knowledge distillation framework that leverages multiple open-source LLMs. The framework first evaluates candidate LLMs to select the best-performing model, which generates a synthetic supervision corpus. A lightweight model is then fine-tuned using QLoRA to produce accurate, hallucination-resistant clinical explanations. As computed in Eq.(16) [19] [21]:

$$E = \{V_{ET}, V_{ED}, V_{NCR}, A, \hat{y}, R, U, \alpha_1, \dots, \alpha_K\}, \quad (16)$$

where V_{ET}, V_{ED}, V_{NCR} denote the habitat-specific tumor volumes, A represents patient age, $\hat{y}$ is the predicted survival time, R is the expected risk score, U is the predictive uncertainty, and α_K represents the prototype assignment weights. The extracted evidence is deterministically transformed into a structured JSON payload, $J = \Phi(E)$, ensuring the reasoning process is strictly grounded in model-derived facts. A prompt template $P = \psi(J)$ instructs the LLM to Generate a concise clinical explanation for the predicted survival outcome. Use only the provided evidence. A synthetic survival explanation dataset is generated for the entire cohort. For each patient i , the input consists of the structured evidence payload J_i , and the output

is generated reasoning report T_i . This process yields a comprehensive corpus of evidence-explanation training pairs is illustrated as Eq. (17) [42] where D denotes the synthetic evidence-explanation dataset.

$$D = \{(J_i, T_i)\}_{i=1}^N \quad (17)$$

This corpus serves as the high-fidelity ground-truth required for downstream instruction tuning. To enable efficient clinical deployment and fast inference, we select the Meta-Llama-3.1-8B-Instruct model. The model is fine-tuned on the synthetic corpus D using Quantized Low-Rank Adaptation (QLoRA). The fine-tuning objective minimizes the causal language modeling loss.

$$L_{tune} = - \sum_{i=1}^N \sum_{t=1}^{|T|} \log P(T_{i,t} | T_{i,<t}, J_i, \phi_s) \quad (18)$$

Pseudo Code: Training & Inference of ProtoSurv-X

Input: Multi-modal BraTS MRI volumes $X = \{T = T1, T1_{ce}, T2, FLAIR\}$ patient age A , survival labels y , prototypes $K = 10$ Survival prediction $\hat{y}$, risk score R , uncertainty estimate U , explanation T .

Phase 1: Segmentation Pre-training each training subject Apply diffusion enhancement to obtain $\hat{x}$ and predict habitat maps $P = \{P_{ET}, P_{NCR}, P_{ED}\}$ Optimize $L_{seg} = \lambda_{DICE} L_{DICE} + \lambda_{CE} L_{CE}$ Freeze segmentation parameters.

Phase 2: Prototype-guided Survival Learning Initialize prototype memory bank $M = \{m_1, m_2, \dots, m_K\}$ each mini-batch Extract and fuse features: $F_{concat} = f_{rad} \oplus f_{deep} \oplus f_{hab} \oplus f_{clin}$ Compute prototype similarities $s_k = \cos(F, m_k)$ and weights $\alpha_k = \text{softmax}(s_k)$ Construct embeddings: $Z = \sum_{k=1}^K \alpha_k m_k$ and $H = W_h[F|Z] + b_h$ predict evidence $e = f_\phi(H)$ and Dirichlet parameters $\alpha = e + 1$ Estimate risk probability p_c , risk score R and uncertainty $U = C / \sum_c \alpha_c$ Predict survival time $\hat{y} = g_\psi(H)$ Compute $L_{total} = L_{surv} + \lambda_1 L_{SCL} + \lambda_2 L_{EDL} + \lambda_3 L_{proto}$ Update network parameters and prototype memory bank M .

Phase 3: Inference and Explanation Generation each test subject Perform forward propagation to obtain $\{P = \{P_{ET}, P_{NCR}, P_{ED}\}, \hat{y}, U, \alpha_k\}$ Construct evidence payload J and generate explanation T using QLoRA-tuned Meta Llama-3.1-8B $\hat{y}, R, U, T$

In Eq. (18) [30], where ϕ_s represents the LoRA adapted parameters and $T_{i,t}$ denotes the token at position t and $T_{i,<t}$, denotes the preceding tokens. This step finalizes the fully open-source clinical reasoning generator, producing a lightweight, highly accurate model tailored specifically for BraTS survival explanations.

G. Model Training Strategy

The proposed framework is optimized in three sequential stages to decouple tumor segmentation from survival prediction. **Stage 1:** The Diffusion-Enhanced SwinUNETR is trained for tumor segmentation using a hybrid Dice and Cross-Entropy loss Eq. (19) [7].

$$L_{seg} = \lambda_{Dice} L_{Dice} + \lambda_{CE} L_{CE} \quad (19)$$

where λ_{Dice} and λ_{CE} are weighting coefficients that balance region overlap accuracy and voxel-wise classification performance, L_{Dice} and L_{CE} denote the Dice and cross-entropy losses, respectively. After convergence, the segmentation network is frozen to generate tumor probability maps and deep features, from which habitat-specific radiomic, habitat, and clinical features are extracted and fused into the patient representation. The prototype learning, survival-aware contrastive learning, and evidential prediction modules are jointly optimized. The survival regression objective is defined as the Mean Absolute Error (MAE)

$$L_{surv} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (20)$$

In Eq. (20) [43] where N denotes the number of patients in the batch, y_i denotes the actual survival time and $\hat{y}_i$ denotes the predicted survival time. The overall training objective is then defined as a multi-task loss function, as illustrated in Eq.(21) [31] [34] [4].

$$L_{total} = L_{surv} + \lambda_1 L_{SCL} + \lambda_2 L_{EDL} + \lambda_3 L_{proto} \quad (21)$$

where λ_1, λ_2 and λ_3 are balancing hyperparameters. L_{SCL} is the survival-aware contrastive loss, L_{EDL} represents the evidential uncertainty learning objective, and L_{proto} corresponds to the prototype memory optimization loss. The complete training and inference workflow of ProtoSurv-X is summarized in the pseudocode shown below.

IV. Experimental Setup and Evaluation

Experiments were conducted on the BraTS 2019 and BraTS 2020 datasets and evaluated in three phases: independent tumor segmentation, survival prediction under five evaluation protocols (including in-domain, cross-cohort, and unified 5-fold cross-validation), and zero-shot evaluation of seven open-source LLMs for model selection.

A. Implementation Details

The framework was implemented in PyTorch using the MONAI framework and trained on two NVIDIA H100 NVL (95 GB) GPUs. During segmentation training, MRI volumes were cropped into $128 \times 128 \times 128$ patches, and the AdamW optimizer was used. The hyperparameters were selected based on validation performance. Subject-level separation was strictly

maintained across all experiments, and survival prediction results are reported as the mean over 5-fold cross-validation. Within each fold, prototype construction and survival model training were performed using only the corresponding training partition, while the held-out fold was used exclusively for evaluation. This fold-wise strategy ensured strict patient-level separation and prevented information leakage. For the clinical reasoning module, deterministic JSON evidence was provided to the LLMs using a temperature of 0.2 and a top-p value of 0.9.

B. Dataset Integrity and Evaluation Protocol: The unified cohort was constructed by merging the BraTS 2019 and BraTS 2020 datasets and removing duplicate subjects using official patient identifiers. After deduplication, 369 unique subjects remained, including 118 GTR patients with complete survival annotations eligible for survival prediction. The dataset was used for four stages: tumor segmentation on all 369 subjects; survival prediction and risk modeling on 118 GTR patients; zero-shot LLM benchmarking on 100 patients sampled exclusively from training partitions; and QLoRA fine-tuning using approximately 94 synthetic explanation samples per fold, generated exclusively from the corresponding training partitions. Survival prediction used patient-level five-fold cross-validation, with test-fold sizes of 24, 24, 24, 23, and 23 patients. The survival cohort comprised 54 short-term (<300 days), 35 intermediate (300–450 days), and 29 long-term (>450 days) survivors, with a median survival of 375 days (IQR: 245–560 days) and 0% censoring. To prevent information leakage, preprocessing, prototype construction, habitat extraction, and model training were performed independently within each fold using only the training partition.

C. Survival Analysis and Risk Stratification

All 118 GTR patients had fully observed overall survival times, resulting in 0% censoring; survival prediction was therefore formulated as continuous regression. For clinical interpretation, predictions were stratified into short-term (<300 days), intermediate (300–450 days), and long-term (>450 days) groups following the established BraTS protocol [23], [24]. These thresholds were used only for post-hoc risk stratification and did not constrain the regression. The C-index was calculated using predicted survival times to assess concordance with observed survival ordering. For IBS, time-dependent survival probabilities were obtained from the model's predicted risk representation and integrated over the evaluation horizon. Kaplan–Meier curves and log-rank tests were used to assess risk-group separation, while calibration curves compared predicted and observed probabilities.

Table 1. Quantitative Survival Prediction Performance on the Unified BraTS Cohort (Five Fold Cross Validation)

Method	MAE (Days) ↓	RMSE (Days) ↓	C-Index ↑	IBS ↓
Cox Proportional Hazards (CPH) [37]	254.8 ± 16.5	312.4 ± 19.8	0.584 ± 0.018	0.231 ± 0.014
Radiomics + RF [2]	238.4 ± 15.2	295.6 ± 18.4	0.612 ± 0.015	0.215 ± 0.012
Random Survival Forests (RSF) [38]	224.6 ± 14.1	281.2 ± 16.5	0.638 ± 0.014	0.201 ± 0.011
Deep Features (SwinUNETR) [8]	205.1 ± 12.8	258.3 ± 14.1	0.654 ± 0.012	0.182 ± 0.009
DeepSurv [39]	192.3 ± 11.5	246.7 ± 13.2	0.671 ± 0.011	0.174 ± 0.010
DeepHit [40]	185.7 ± 10.8	239.5 ± 12.4	0.684 ± 0.011	0.166 ± 0.009
CGNN [1]	178.2 ± 9.9	231.4 ± 11.8	0.692 ± 0.010	0.161 ± 0.008
MTC-Net [14]	172.5 ± 9.4	224.8 ± 11.2	0.698 ± 0.010	0.158 ± 0.008
DPSurv (Adapted) [4]	161.4 ± 8.5	210.6 ± 10.1	0.718 ± 0.009	0.142 ± 0.007
Proposed ProtoSurv-X	145.2 ± 6.5	189.4 ± 8.3	0.745 ± 0.008	0.124 ± 0.005

Table 2. Survival Prediction Performance Across the Five in-domain, cross-cohort, and unified BraTS evaluation protocol

Protocol	Evaluation Regime	MAE (Days) ↓	RMSE (Days) ↓	C-Index ↑
Experiment 1	BraTS 2019 Baseline (In-Domain)	158.4 ± 8.1	205.1 ± 11.4	0.720 ± 0.012
Experiment 2	BraTS 2020 Baseline (In-Domain)	152.6 ± 7.5	198.5 ± 10.2	0.731 ± 0.010
Experiment 3	Forward Generalization (19 → 20)	168.3 ± 9.6	216.7 ± 12.8	0.695 ± 0.014
Experiment 4	Reversed Generalization (20 → 19)	164.8 ± 8.9	212.3 ± 11.5	0.704 ± 0.013
Experiment 5	Unified Cohort (5-Fold CV)	145.2 ± 6.5	189.4 ± 8.3	0.745 ± 0.008

V. Results

The unified survival cohort was constructed by integrating the BraTS 2019 and BraTS 2020 datasets following the official BraTS survival prediction protocol. After applying the GTR-based inclusion criteria and removing duplicate subjects, the final cohort comprised 369 patients, including 118 GTR cases with complete survival data.

A. Cohort Harmonization and Deduplication

Since BraTS 2020 is an extension of the BraTS 2019 challenge, both releases contain overlapping subjects. To prevent duplicate samples and information leakage, patient records were matched using the official BraTS patient identifiers provided with the datasets. Subjects appearing in both releases were retained only once in the unified cohort, while duplicate entries from the earlier release were removed. This identifier-based verification ensured that each patient contributed a single MRI examination and survival record to the final dataset.

B. Brain Tumor Segmentation Results

The structural robustness of the proposed Diffusion-Enhanced SwinUNETR was evaluated independently on the BraTS 2019 [2] and 2020 [23] datasets against established baselines (3D U-Net, nnU-Net, TransBTS, and SwinUNETR). On BraTS 2019, the proposed framework achieved state-of-the-art Dice Similarity Coefficients (DSC) of 91.24%, 87.38%, and 81.76% for the Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET), respectively. Notably, the diffusion-based enhancement significantly reduced the 95th percentile Hausdorff Distance (HD95) for the highly irregular ET to 2.84 mm. This superiority was maintained on the expanded BraTS 2020 cohort (WT DSC: 91.82 %, ET DSC: 82.75%), confirming that continuous probability mapping successfully mitigates center-specific imaging protocols and provides an anatomically precise foundation for downstream survival analysis.

C. Survival Prediction and Generalization

The survival prediction performance was evaluated on the unified BraTS cohort against traditional, standard deep learning, and state-of-the-art baselines. As shown in Table 1, the proposed **ProtoSurv-X** consistently outperforms all traditional statistical, deep survival, and

Table 3. Comprehensive Ablation Study on Framework Components

Model Configuration	MAE (Days) ↓	C-Index ↑
Baseline (Direct MLP, No Prototypes)	168.4 ± 10.2	0.690 ± 0.012
+ Prototype Learning (Only)	159.1 ± 8.6	0.712 ± 0.010
+ Prototype & Evidential (Without SCL)	156.8 ± 7.9	0.722 ± 0.011
+ Prototype & SCL (Without Evidential)	150.3 ± 7.4	0.734 ± 0.009
Full Model (Proto + SCL + Evidential)	145.2 ± 6.5	0.745 ± 0.008

Table 4. Extended component removal ablation study of the ProtoSurv-X architecture.

Model Configuration	MAE (Days) ↓	C-Index ↑
Full ProtoSurv-X Framework	145.2 ± 6.5	0.745 ± 0.008
Without diffusion enhancement	153.8 ± 7.2	0.729 ± 0.010
Without Probabilistic Habitats (using Hard Masks)	157.3 ± 8.1	0.721 ± 0.011
Without Evidential Uncertainty Loss	151.4 ± 7.6	0.732 ± 0.009
Clinical Age	148.9 ± 6.8	0.739 ± 0.0

Table 5. Feature-modality ablation analysis for glioma survival prediction

Input Features	MAE (Days) ↓	C-Index ↑
Age	285.4 ± 18.5	0.550 ± 0.018
Radiomics Only	240.1 ± 14.8	0.608 ± 0.015
Habitat Features	225.3 ± 13.2	0.625 ± 0.014
Deep Features	210.5 ± 11.5	0.645 ± 0.012
Rad + Deep	178.6 ± 9.8	0.695 ± 0.011
Rad + Deep + Habitat	158.4 ± 8.2	0.728 ± 0.009
Full Model (All Features)	145.2 ± 6.5	0.745 ± 0.008

multimodal baseline methods on the unified BraTS cohort. While conventional linear and tree-based models such as Cox Proportional Hazards (CPH) and Random Survival Forests (RSF) struggle to capture complex non-linear imaging biomarkers (yielding C-Indices of 0.584 and 0.638, respectively), deep survival architectures like DeepSurv and DeepHit improve prognostic ranking by leveraging dense feature representations. Compared with recent multimodal and joint-learning frameworks including Cooperative GNN [1] and MTC-Net [14], our ProtoSurv-X substantially reduces the Mean Absolute Error (MAE) from 172.5 to 145.2 days and increases the C-Index from 0.698 to 0.745. Furthermore, while an MRI-adapted implementation of the evidential DPSurv [4] model achieves a competitive ranking (C-Index: 0.718), our approach delivers superior overall calibration, as shown by the lowest Integrated Brier Score (IBS) of 0.124. Wilcoxon signed-rank testing on paired five-fold cross-validation predictions confirmed that these improvements in MAE, C-Index, and IBS over all evaluated baselines are statistically significant($p<0.05$). Table 2 demonstrates the robustness of the proposed framework across in-domain and cross-cohort evaluations. Despite distribution shifts, the model achieves competitive performance in both forward (C-Index: 0.695) and

reverse (0.704) generalization, indicating strong robustness. The unified cohort (Exp 5) achieves the best overall performance, validating the effectiveness of cohort deduplication and unified training.

D. Analysis of Component Contributions

As shown in Table 3, prototype learning substantially improves survival prediction over direct regression (C-Index: 0.690 → 0.712), while incorporating survival-aware contrastive learning (SCL) further increases the C-Index to 0.745 by learning more discriminative prognostic representations. The complete framework, integrating prototype learning, SCL, and evidential uncertainty estimation, achieves the best overall performance across all evaluation metrics. Prototype learning organizes heterogeneous patients into recurring prognostic patterns, while survival-aware contrastive learning encourages survival-consistent representations, providing complementary gains beyond direct regression. The extended ablation shows in Table 4, that removing diffusion enhancement and probabilistic habitats reduces performance (C-index: 0.729 and 0.721), while removing evidential loss and age causes smaller declines (0.732 and 0.739), supporting their complementary contributions.

VI. Discussion

1. Feature and Habitat Contribution Analysis

Table 6:Zero-shot evaluation of candidate LLMs for synthetic clinical-explanation corpus generation.

CandidateLLM	BERTScore↑	Hallucination R (%) ↓
Mistral-7B[24]	0.842	8.1%
Llama-3.1-8B-I [25]	0.852	7.5%
Phi-4 [26]	0.865	6.4%
Gemma-2-27B [27]	0.878	5.2%
Qwen2.5-72B[28]	0.905	3.8%
Llama-3.3-70B [25]	0.912	3.5%
DeepSeek-R1[29]	0.924	2.1%

Table 7. Effect of QLoRA fine tuning on the performance of the light weight Llama3.1

Model Configuration	BERTScore ↑	Hallucination Rate (%) ↓
DeepSeek-R1-32B [29]	0.924	2.1%
Llama-3.1-8B (BeforeQLoRA) [25]	0.852	7.5%
Llama-3.1-8B (After QLoRA, Proposed)	0.915	2.6%

While the previous ablation study demonstrated the contribution of the proposed architectural components, this experiment investigates the individual prognostic significance of different input feature modalities. Specifically, we evaluate the impact of clinical, radiomic, deep imaging, and habitat-specific features on survival prediction through a comprehensive feature ablation study, as summarized in Table 5.

1. Feature Modality Ablation
Evaluated independently, clinical demographic data (Age) yields the weakest prognostic power (C-Index: 0.550), while Deep spatial embeddings outperform standard Radiomics (0.645 vs. 0.608). However, progressive fusion demonstrates clear complementary value. Combining Radiomics, Deep Features, and Habitat metrics incrementally improves the C-Index to 0.728. The Full Model achieves optimal performance (MAE: 145.2, C-Index: 0.745), confirming that no single modality captures the full spectrum of glioma survival dynamics.

2. Habitat Importance:

To determine which specific spatial regions drive these predictions, a permutation-based feature importance analysis was applied to the extracted habitat stream.

F. Evidence-Grounded Explanation Generation
Seven open-source LLMs were evaluated using structured evidence from 100 validation patients, with BERTScore and hallucination rate used as evaluation metrics. As shown in Table 6, the models exhibited different levels of linguistic similarity and factual consistency. BERTScore increased from 0.842 for Mistral-7B-Instruct-v0.3 to 0.924 for DeepSeek-R1 Distill-Qwen-32B, while the hallucination rate decreased from 8.1% to 2.1%. The intermediate models, including Meta-Llama-3.1-8B-Instruct, Phi-4, Gemma-2-27B-IT, Qwen2.5-72B-Instruct, and Llama-3.3-70B-Instruct, showed intermediate performance. These results are consistent with previous applications of open-source LLMs in specialised medical tasks, such as structured medical report generation using Llama-based models [25] and MRI-report processing using Qwen2.5 [28]. However, the present evaluation specifically considers evidence-grounded prognostic explanation generation using quantitative outputs from

Table 8. Evidence-grounded clinical explanations and predicted survival-risk categories for the three representative cases.

Cases	Generated Explanation(Clinical Interpretation)	Predicted Survival Risk Category
Case1	Large enhancing core (H2) and peritumoral edema (H3) dominate, indicating aggressive tumour behaviour and poor prognosis..	High Risk; Predicted Survival: 11.2 Months (342 Days); Risk Score: 0.86; Confidence: 0.91
Case 2	Moderate edema (H3) and enhancing regions (H2) show a mixed habitat pattern, indicating intermediate aggressiveness and moderate survival	Medium Risk; Predicted Survival: 18.7 Months (570 Days); Risk Score: 0.54; Confidence: 0.78
Case 3	Predominant infiltrative region (H3) and edema (H3), with a smaller enhancing core (H2), indicate lower high-risk habitat proportion and longer expected survival.	Low Risk; Predicted Survival: 28.3 Months (862 Days); Risk Score: 0.21; Confidence: 0.88

the survival prediction model. Based on the results, DeepSeek-R1-Distill-Qwen-32B was selected as the teacher model for generating the synthetic clinical-explanation corpus [29]. To evaluate knowledge transfer to a smaller model, Meta-Llama-3.1-8B-Instruct was fine-tuned using QLoRA on the synthetic corpus generated by DeepSeek-R1-Distill-Qwen-32B. Fig. 2 illustrates qualitative visualization of multimodal brain MRI sequences, tumour segmentation, habitat maps, and survival-risk heatmaps for three representative cases. FLAIR, T1, T1ce, and T2 denote the MRI sequences used for tumour analysis. As shown in Table 7, BERTScore increased from 0.852 to 0.915, while the hallucination rate decreased from 7.5% to 2.6%. The fine-tuned 8B model therefore approached the teacher model performance of 0.924 BERTScore and 2.1% hallucination. This improvement is consistent with the parameter-efficient adaptation objective of QLoRA [30]. Unlike general-purpose fine-tuning, the present approach uses structured prognostic evidence as the supervision source, thereby constraining the model to generate explanations from

model-derived survival information. These results demonstrate the effectiveness of QLoRA for adapting a lightweight LLM to the specific task of evidence-grounded prognostic explanation generation. The fine-tuned version achieves a BERTScore of 0.915 and dramatically reduces its hallucination rate to just 2.6%, proving that complex prognostic reasoning can be successfully transferred into a highly efficient, deployable 8B-parameter model.

G. Hallucination-Identification Protocol

BERTScore evaluates linguistic fidelity, while hallucination rate was assessed using a deterministic, rule-based protocol against the source JSON payload. Fig 2 presents the MRI images for Cases 1, 2, and 3, while the corresponding explanation-generation results are presented in Table 8. Explanations were checked for (i) factual consistency of survival days, risk category, and habitat volumes, (ii) explicit communication of high uncertainty, and (iii) absence of unsupported treatment recommendations or clinical claims. Large-scale blinded neuro-oncologist

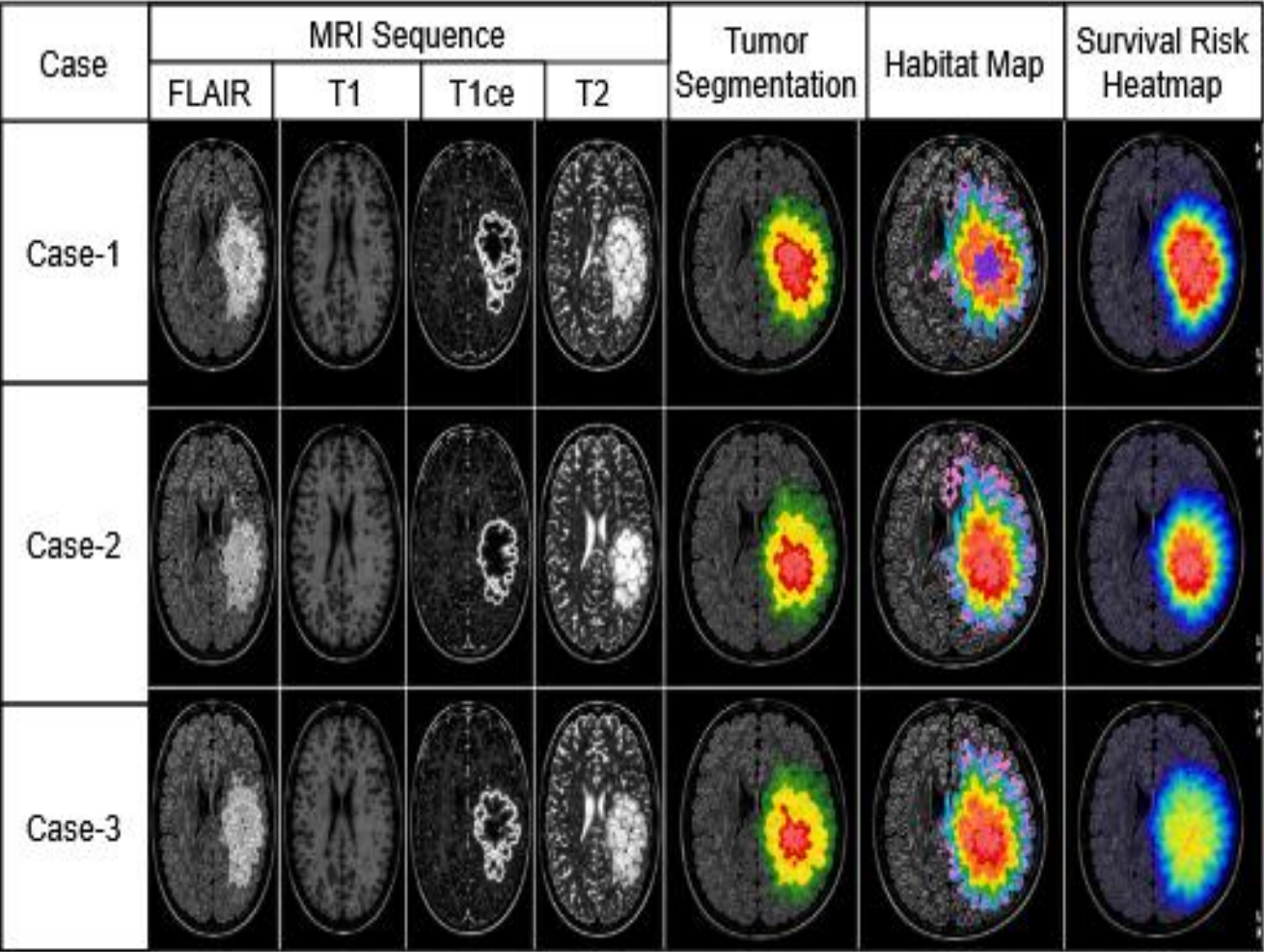


Fig 2. Qualitative visualisation of multimodal brain MRI sequences.

evaluation remains a future validation step because ProtoSurv-X is currently a computational proof-of-concept. To clarify ProtoSurv-X's novelty, we compared it with MTC-Net, CGNN, MMP, and DPSurv. While MTC-Net and CGNN use global MRI representations, MMP introduces prototype learning and DPSurv incorporates evidential prediction; none combines 3D probabilistic habitats, multimodal prototype-guided learning, evidential uncertainty, and evidence-grounded LLM explanations. ProtoSurv-X unifies these components into a single prognostic framework.

Despite promising results, ProtoSurv-X has several limitations. First, the relatively small cohort of 118 GTR patients may limit statistical power and increase cohort-specific effects. Second, evaluation on only the BraTS 2019–2020 benchmark limits validation across independent clinical populations and acquisition settings. Third, the retrospective public data may introduce selection and dataset-specific biases. Fourth, reliance on MRI and age excludes molecular and genetic biomarkers that may provide complementary prognostic information. Fifth, the static prototype memory may not adapt to evolving multi-center populations. Finally, the LLM explanations rely on synthetic evidence–explanation pairs and require validation by neuro-oncology experts and prospective multi-center studies to establish clinical reliability, generalizability, and safe deployment in live clinical workflows.

VII. Conclusion

This study developed ProtoSurv-X, an explainable framework for MRI-based glioma survival prediction integrating tumor segmentation, habitat modeling, prototype-guided learning, uncertainty estimation, and evidence-grounded LLM explanations. The framework achieved 91.24%, 87.38%, and 81.76% Dice for WT, TC, and ET, respectively, with a C-index of 0.745, a BERTScore of 0.915, and a hallucination rate of 2.6%. These results demonstrate ProtoSurv-X's potential for accurate, uncertainty-aware, and explainable prognostic decision support. Future work will focus on larger multi-center cohorts, external validation, molecular and genetic biomarkers, adaptive prototype memory, and neuro-oncology expert evaluation to improve robustness, generalizability, and clinical safety.

Acknowledgment

The authors acknowledge the institutions and publicly available BraTS datasets that supported this research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets analyzed in this study are publicly available from the BraTS 2019 and BraTS 2020 repositories. No new patient data were collected or generated as part of this study. The datasets used are available from the sources specified in the manuscript.

Author Contribution

Author Contributions: Neeshu Kumari: Conceptualization, methodology, software, investigation, data analysis, visualization, and writing original draft. Saurabh Shah: Supervision and writing review and editing. Chintan Thacker: Validation, formal analysis, investigation, and writing review and editing. All authors reviewed and approved the final manuscript.

Declarations

Ethical Approval

Ethical approval was not required because this study used only publicly available, de-identified datasets from the BraTS 2019 and BraTS 2020 challenges and involved no direct recruitment, intervention, or collection of identifiable patient information

Consent for Publication Participants.

Consent for Publication: Not applicable. No identifiable personal information, images, or data from individually identifiable participants are reported in this manuscript.

Competing Interests

The authors declare that they have no competing financial or non-financial interests that could have influenced the work reported in this manuscript.

Declaration of Generative AI and AI-Assisted Technologies:

Large language models were used as components of the proposed research framework for generation and evaluation of evidence-grounded clinical explanations, as described in the Methods section. The authors remain fully responsible for the scientific content, interpretation, accuracy, and integrity of the manuscript.

References

- [1] R. Cao, X. Hu, L. Xiao, G. Qu, H. Huo, V. D. Calhoun, Y.-P. Wang, and X. Sun, "Cooperative multiplex GNN for high-grade glioma survival prediction from preoperative multi-modal radiomics-based brain networks," *IEEE Trans. Med. Imaging*, vol. 45, no. 7, pp. 3464–3476, 2026.
<https://doi.org/10.1109/TMI.2026.3677065>
- [2] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara et al.,

- "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," in *BrainLesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2019, pp. 374–387. <https://doi.org/10.17863/CAM.38755>
- [3] A. H. Song, R. J. Chen, G. Jaume, A. J. Vaidya, A. S. Baras, and F. Mahmood, "Multimodal prototyping for cancer survival prediction," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024, pp. 46050–46073. <https://dl.acm.org/doi/10.5555/3780338.3781154>
- [4] Y. Xing, L. Huang, J. Ma, R. Hong, J. Qiu, P. Liu, K. He, H. Fu, and M. Feng, "DPSurv: Dual-prototype evidential fusion for uncertainty-aware and interpretable whole slide image survival prediction," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2026. <https://icml.cc/virtual/2026/papers.html>
- [5] M. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Cham, Switzerland: Springer, 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
- [6] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vision (3DV)*, Stanford, CA, USA, 2016, pp. 565–571. <https://doi.org/10.1109/3DV.2016.79>
- [7] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021. <https://doi.org/10.1038/s41592-020-01008-z>
- [8] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI BrainLes Workshop*, Cham, Switzerland: Springer, 2021, pp. 272–284. https://doi.org/10.1007/978-3-031-08999-2_22
- [9] H. Liu, D. Wei, D. Lu, J. Sun, L. Wang, and Y. Zheng, "M3AE: Multimodal representation learning for brain tumor segmentation with missing modalities," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, pp. 1657–1665, 2023. <https://doi.org/10.1609/aaai.v37i2.25253>
- [10] G. Liang, Q. Zhou, Z. Wang, J. Chen, L. Gu, C. Yao, S. Wu, B. Huang, and K. Chen, "Semantic-guided masked mutual learning for multi-modal brain tumor segmentation with arbitrary missing modalities," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 5, pp. 5137–5145, 2025. <https://doi.org/10.1609/aaai.v39i5.32545>
- [11] Y. Zhao, C. Chen, Q. Y. Pang, Y. Fu, Q. Li, C. Tang, B. T. Ang, and Y. Jin, "Tackling dual-stage missing modalities in brain tumor segmentation via robust modality reconstruction and prompt-guided modality adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 16, pp. 13314–13322, 2026. <https://doi.org/10.1609/aaai.v40i16.38334>
- [12] T. Liu, H. Jiang, and K. Huang, "KMD: Koopman multi-modality decomposition for generalized brain tumor segmentation under incomplete modalities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 15663–15671, 2025. <https://doi.org/10.1109/CVPR52734.2025.01460>
- [13] H. Yang, J. Sun, and Z. Xu, "Learning unified hyper-network for multi-modal MR image synthesis and tumor segmentation with missing modalities," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3678–3689, 2023. <https://doi.org/10.1109/TMI.2023.3301934>
- [14] I. Mazumdar and J. Mukhopadhyay, "Multi-scale transformer-CNN network for brain tumor segmentation and survival prediction," in *Proc. 15th Indian Conf. Comput. Vis. Graph. Image Process. (ICVGIP)*, 2024, pp. 1–9. <https://doi.org/10.1145/3702250.3702257>
- [15] A. Datta, S. Sarkar, A. L. Sharma, and P. Ghosal, "Multimodal exponential moving average-guided representation for integrated tumor-segmentation and survival prediction," *Eng. Appl. Artif. Intell.*, vol. 167, Art. no. 113868, 2026. <https://doi.org/10.1016/j.engappai.2026.113868>
- [16] L. Qu, J. Xiao, X. Liu, C. Sun, H. Cui, Y. Fang, R. Su, Q. Jin, and L. Wei, "MDCS-MoAME: Multi-directional composite scanning with mixture of attention and Mamba experts for cancer survival prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026, pp. 14461–14470. https://openaccess.thecvf.com/content/CVPR2026/html/Qu_MDCS-MoAME_Multi-directional_Composite_Scanning_with_Mixture_of_Attention_and_Mamba_CVPR_2026_paper.html
- [17] A. H. Song, R. J. Chen, T. Ding, D. F. K. Williamson, G. Jaume, and F. Mahmood,

- "Morphological prototyping for unsupervised slide representation learning in computational pathology," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 11566–11578.
<https://doi.org/10.1109/CVPR52733.2024.01099>
- [18] J. Wu, M. Chen, X. Ke, T. Xun, X. Jiang, H. Zhou, L. Shao, and Y. Kong, "Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 5144–5153.
<https://doi.org/10.1109/CVPR52734.2025.00485>
- [19] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems*, vol. 36, pp. 28541–28564, 2023.
<https://doi.org/10.52202/075280-1240>
- [20] C. Wu, X. Zhang, Y. Zhang, H. Hui, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology by leveraging web-scale 2D and 3D medical data," *Nat. Commun.*, vol. 16, no. 1, Art. no. 7866, 2025. <https://doi.org/10.1038/s41467-025-62385-7>
- [21] B. Kandala and R. K. G. V. S. Raj Kumar, "Trustworthiness Index SHapley Additive exPlanations-based skin cancer diagnosis via hybrid deep concept bottleneck learning and LLM-driven reasoning," *Biomed. Signal Process. Control*, vol. 122, Art. no. 110309, 2026.
<https://doi.org/10.1016/j.bspc.2026.110309>
- [22] B. W. Patterson et al., "Evaluating clinical AI summaries with large language models as judges," *npj Digital Medicine*, vol. 8, Art. no. 640, 2025. <https://doi.org/10.1038/s41746-025-02005-2>
- [23] T. Henry, A. Carré, M. Lerousseau, T. Estienne, C. Robert, N. Paragios, and E. Deutsch, "Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-Net neural networks: A BraTS 2020 challenge solution," in *BrainLesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Cham, Switzerland: Springer, 2021, pp. 327–339.
https://doi.org/10.1007/978-3-030-72084-1_30
- [24] Samo, H., Ali, K., Memon, M., Abbasi, F.A., Koondhar, M.Y. and Dahri, K., 2024. Fine-tuning mistral 7b large language model for python query response and code generation: A parameter efficient approach. *VAWKUM Transactions on Computer Sciences*, 12(1), pp.205-217.
<https://doi.org/10.21015/vtcs.v12i1.1885>
- [25] Prucker, P., Bressemer, K.K., Kim, S.H., Weller, D., Kader, A., Dorfner, F.J., Ziegelmayer, S., Graf, M.M., Lemke, T., Gassert, F. and Can, E., 2026. Privacy-preserving generation of structured lymphoma progression reports from cross-sectional imaging: A comparative analysis of Llama 3.3 and Llama 4. *Journal of Imaging Informatics in Medicine*, 39(2), pp.1868-1878.
<https://doi.org/10.1007/s10278-025-01618-z>
- [26] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang, "Phi-4 Technical Report," *Microsoft Research*, MSR-TR-2024-57, 2024.
<https://www.microsoft.com/en-us/research/publication/phi-4-technical-report/>
- [27] T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramár, A. Dragan, R. Shah, and N. Nanda, "Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2," in *Proc. 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024, pp. 278–300.
<https://doi.org/10.18653/v1/2024.blackboxnlp-1.19>
- [28] K. A. Huang, S. V. Prakash, D. Samvelian, N. S. Prakash, and S. Prakash, "Implementation of a locally deployed Qwen2.5-7B-Instruct pipeline for structured classification of elbow MRI reports," *Cureus*, vol. 18, no. 8, 2026.
<https://doi.org/10.7759/cureus.114918>
- [29] D. Guo et al., "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, vol. 645, pp. 633–638, 2025.
<https://doi.org/10.1038/s41586-025-09422-z>
- [30] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Advances in Neural Information Processing Systems*, vol. 36, pp. 10088–10115, 2023.
<https://doi.org/10.52202/075280-0441>
- [31] Z. Li, W. Chen, H. Zhong, and C. Liang, "PCLSurv: A prototypical contrastive learning-based multi-omics data integration model for cancer survival prediction," *Briefings in Bioinformatics*, vol. 26, no. 2, Art. no. bbaf124, 2025.
<https://doi.org/10.1093/bib/bbaf124>
- [32] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int.*

- Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Cham, Switzerland: Springer, 2021, pp. 109–119. https://doi.org/10.1007/978-3-030-87193-2_11.
- [33] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, vol. 119, 2020, pp. 1597–1607. <https://proceedings.mlr.press/v119/chen20j.html>
- [34] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 18661–18673. https://proceedings.neurips.cc/paper_files/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html
- [35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. <https://doi.org/10.5555/3495724.3496298>
- [36] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 8162–8171. Available: <https://proceedings.mlr.press/v139/nichol21a.html>
- [37] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc. Ser. B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- [38] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, 2008. <https://doi.org/10.1214/08-AOAS169>
- [39] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, no. 1, Art. no. 24, 2018. <https://doi.org/10.1186/s12874-018-0482-1>
- [40] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "DeepHit: A deep learning approach to survival analysis with competing risks," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, New Orleans, LA, USA, 2018, pp. 2314–2321. <https://doi.org/10.1609/aaai.v32i1.11842>
- [41] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 3179–3189, 2018. <https://doi.org/10.5555/3327144.3327239>
- [42] C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister, "Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes," in *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023. <https://doi.org/10.18653/v1/2023.findings-acl.507>
- [43] S.-A. Qi, N. Kumar, M. Farrokh, W. Sun, L.-H. Kuan, R. Ranganath, R. Henao, and R. Greiner, "An Effective Meaningful Way to Evaluate Survival Models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, vol. 202, 2023, pp. 28244–28276. <https://proceedings.mlr.press/v202/qi23b.html>

Author Biography



Neeshu Kumari completed her Bachelor of Engineering in Computer Science and Engineering from Uttar Pradesh Technical University (UPTU) in 2010, and received her M.Tech degree in Computer Science and Engineering from GLA University in 2013. She is currently pursuing a Ph.D. in Computer Science and Engineering from Parul University. She has over 8 years of academic experience. Currently, she serves as an Assistant Professor in the Department of Computer Science and Engineering at GSFC University, and her research interests include machine learning, deep learning, computer vision, and wireless networking. She has guided many major and minor projects of undergraduate students. She is eager to develop digital study materials to provide her students with easy access during their studies.



Prof. Saurabh Shah is a Professor of Computer Science & Engineering and a passionate Academician with more than 24 Years of experience in Academics, Administration and Research. He is presently working as a Professor, Director (Talent and Career Development) and Director (Faculty Augmentation) at Parul University, Gujarat, India. His major Research areas are Medical Image Analysis, Computer Vision, Data Analytics, Machine Learning and Pattern Recognition. He has supervised 08 PhD candidates who have been awarded PhD under his supervision.



Dr. Chintan Thacker received his Ph.D. in Artificial Intelligence & Computer Vision from Gujarat Technological University in 2021. He previously served as the Head of the Department of Computer Science and Engineering at HJD Institute of

Technical Education and Research, Kera, India and as an Associate Professor in the Computer Science and Engineering Department at Parul University, Vadodara, Gujarat. He has over 15 years of academic experience and more than 1 year of industry experience. Currently, he is an Assistant Professor in the Department of Computer Science at KSKV Kachchh University, Gujarat. He has mentored several Ph.D. students & received grants from SSIP & GUJCOST. He regularly conducts workshops and leads projects in the field of computer vision and AI.