Interpreting Survival Predictor Model for Glioblastoma using Explainable Artificial Intelligence

Snehal Rajput[1][0000−0001−8240−3740], Rupal A. Kapdi[2][0000−0003−1995−4149], and Mehul S. Raval[3][0000−0002−3895−1448] Mohendra Roy[4][0000−0001−5815−3294]

[1] School of Computer Science Engineering & Technology, Bennett University, U.P, India; snehal.rajput@bennett.edu.in
[2] Institute of Technology, Nirma University, Ahmedabad, Gujarat, India; rupal.kapdi@nirmauni.ac.in
[3] School of Engineering and Applied Science, Ahmedabad University, Ahmedabad, India.
[4] School of Engineering and Applied Science, Ahmedabad University, Ahmedabad, India.
Corresponding authors: mehul.raval@ahduni.edu.in, mohendra.roy@sot.pdpu.ac.in

Abstract. Gliomas, graded as type-IV tumors, are linked to poor prognosis and low survival chances. An accurate survival prediction model aids in strategically planning patients' treatments. We derive a robust feature set for accurate survival days (SD) prediction, including radiomics, location-based features, and age from the triplanar segmentation network. We study features' global and local impact on SD using various post-hoc explainable AI (XAI) methods. However, post-hoc methods can produce different results for SD prediction, raising the question of interpretability. Therefore, we cross-evaluated the results and found these post-hoc XAI methods were consistent in their interpretations, indicating the robustness of the feature set. Additionally, we establish the biological significance of imaging features better to understand their impact on tumor behavior and patient outcomes. Our SD prediction regressor model outperforms current methods. BraTS2020 validation results showed improvements of 37.7% in accuracy, 16.85% in mean squared error, and 85.8% in Spearman's rank correlation compared to the top-ranking model of the BraTS2020 challenge.

Keywords: Brain tumor segmentation · interpretability · radiomics feature · random forest · survival days.

1 Introduction

Automated Brain tumor segmentation (BTS) and accurate prediction of Survival days (SD) for brain tumor patients are among the most critical tasks in medical image processing. Developing a computational model capable of exceeding human-level segmentation and accurately predicting SD would significantly

enhance the capabilities of healthcare professionals. It would improve the precision, reliability, and standardization of disease diagnosis, treatment planning, and monitoring. Gliomas, arising from glial cells, are the most prevalent and highly malignant brain tumors, associated with elevated rates of morbidity, recurrence, and mortality [21]. MRI scans are most widely used to diagnose tumors because of their non-radiation, high resolution, and high contrast among soft tissues. Multimodal BraTS (Brain tumor segmentation) challenge [6, 28] offers MRI images, and the task includes segmenting tumor regions into Enhancing/Active tumor (ET/AT), Tumor core (TC), and Whole tumor (WT). Following brain tumor segmentation, the subsequent step involves predicting the survival days for glioma patients. This prediction is based on extracting manually crafted features from the segmented outcomes. Furthermore, the objective is to classify the predicted SD into one of three categories: short-term survival (for patients with SD $<$ 300 days), mid-term survival (between SD 300 and 450 days), or long-term survival (with SD $>$ 450 days) [28]. Typically, the evaluation of SD prediction performance involves using standard metrics such as accuracy (ACC), Mean squared error (MSE), median squared error (medianSE), and the Spearman rank coefficient (SRC).

1.1 Challenges in Brain Tumor Segmentation and Survival Days Prediction

Tumor cells display significant size, shape, and location heterogeneity, with complex boundary interactions [37]. Moreover, challenges arise from variations in imaging protocols and limited access to annotated data [33], leading to class imbalance due to fewer tumor pixels. Similarly, predicting survival duration is highly challenging, facing several vital factors, including dependency on the BTS performance, limited availability of comprehensive clinical patient data, and qualitative image characteristics obtained from radiographic images. Recent research suggests that radiomic features hold the potential to capture crucial phenotypic details, including intra-tumor heterogeneity, offering valuable insights for personalized therapy [44, 50]. However, their lack of uniform extraction protocol and interpretability often constrain these features' practical usefulness [15, 50]. Moreover, depending on Machine learning (ML) models for predicting BTS and SD is often perceived as a black box due to the difficulty in interpreting decisions, posing a significant challenge. Consequently, there is an increasing need for Explainable AI (XAI) for ML models.

1.2 Recent Developments in BTS and SD Prediction

Recently, convolutional neural networks (CNN) have significantly progressed in medicine. There has been a surge of 2D and 3D UNet-based networks and ensemble methodologies proposed for BTS. For example, 2D DeepSCAN [26], 2.5D ensemble model [34, 35, 41], 3D UNet [10], 3D UNet++ [53], 3D nnUNet [17], 3D Swin UNETR [16], 3D TranBTS [49] have been proposed. Each subsequent

model has shown improved segmentation performance compared to the predecessor model with added training parameters. 3D models and ensembles represent the current leading approaches for tumor segmentation. Despite its remarkable success in segmentation performance, several potential drawbacks exist, such as the need for extensive annotated data, inherent complexity of models, vulnerability to overfitting, interpreting model, and performance limitations. Moreover, the many trainable parameters and high computational complexity present challenges for achieving rapid medical image segmentation in real-time therapy and diagnosis [51]. Therefore, many efficient networks have been introduced to mitigate these limitations, maintaining high performance while addressing constraints. Examples include lightweight UNet [44] with atrous convolution blocks [2], ERV-Net [52], ESPNet [27], and Triplanar network (TN/2.5D) [34, 35, 42].

Similarly, in the context of SD predictions, existing literature indicates that morphological [3, 13, 33], spatial location [8, 36] and radiomics-based features [1, 24, 36] have demonstrated significant importance. In summary, the literature survey demonstrates that efficient and computationally lightweight networks have the potential to achieve state-of-the-art performance for BTS. Furthermore, it emphasizes the prevalence of statistical, shape, texture, and spatial features in predicting the SD of brain tumor patients.

Hence, in this paper, we investigate the potential and consequences of employing computationally efficient networks for BTS in predicting survival days of brain tumor patients. Additionally, we analyze the behavior of features used for SD predictions and aim to establish the biological relevance of the top four imaging features, ensuring alignment with medical insights through post-hoc interpretability tools. For the BTS, we employed triplanar network (TN) networks [35]. In this TN approach, Multiple 2D models are trained on planar views of MRI images (axial, coronal, and sagittal), and their segmentation outcomes are combined to produce a final segmentation map. This approach effectively balances computational efficiency and performance by integrating 3D and 2D networks.

For predicting SD, we applied the methodology outlined in [36] to extract image-based and radiomic-based features commonly found in literature surveys. Further, Permutation importance (PI) and SRC were employed to select the most significant features from the BTS TN networks. Additionally, we assess the reliability and consistency of the feature set derived from these TN networks for predicting SD. By increasing or decreasing SD, we used interpretation tools to investigate how these features affect SD prediction in global (considering all samples) and local (considering a single sample) scenarios. This analysis helps us understand the impact of these features on SD prediction and identify their biological significance with brain tumor characteristics. Post-hoc interpretation methods have recently become essential for explaining ML models [45]. However, different post-hoc interpretation techniques yield varying results for the same task, which raises the question of which method is the most reliable for accurate post-hoc interpretability [45].

Consequently, we cross-validated the visual graphs obtained for the features from standard post-hoc interpretation methods and evaluated the robustness and reliability of the feature set. We employed the BRATS2020 dataset to showcase the effectiveness of our approach.

In summary, the primary contributions of this paper include:

1. Assessing the robustness of the feature set derived from the triplanar segmentation network.
2. Evaluating the regressor model's performance using the BraTS2020 validation dataset [39].
3. Generating visual representations to investigate how features contribute to predicting the survival days of patients in both global and local scenarios.
4. Investigating the biological relevance of features and linking them with medical insights.

The sections of the paper are organized as follows: In Section 2, we examine the data sources and evaluation metrics. The methodology proposed for SD prediction is detailed in Section 3, while Section 4 provides experimental results and discussions. The Section 5 concludes and suggests future work.

2 Data sources

This study employs the BraTS challenge [6, 28] 2020 dataset. The challenge provides a set of multi-modal magnetic resonance imaging (MRI) volumetric images and promotes the development of algorithms to segment brain tumors and predict the SD of patients. The data sources' details can be displayed in Table 1.

Table 1: BraTS2020 challenge dataset details.

No. of samples	BTS (3D MRI Images)	SD prediction (CSV file)
Training Set:	369 samples	236 (include Age, survival days, and GTR status) (where 117 samples have GTR resection).
Modalities:	T1-weighted post-contrast (T1Gd), T1, T2-weighted (T2), Fluid Attenuated Inversion Recovery (FLAIR) and manual ground truth segmentation	NA
Dimensions:	$155 \times 240 \times 240$ (Depth × Width × Height)	NA
Annotations:	Label 0 for background pixels, 1 for NET/NCR, 2 for ED, 4 for ET, and 0	NA
Validation Set*	125 samples	29 samples (GTR status)
Test Set**	166 samples	107 samples (GTR resection status)

* Ground truth labels are not accessible to the public.
** Test sets are only available to participants in the challenge.

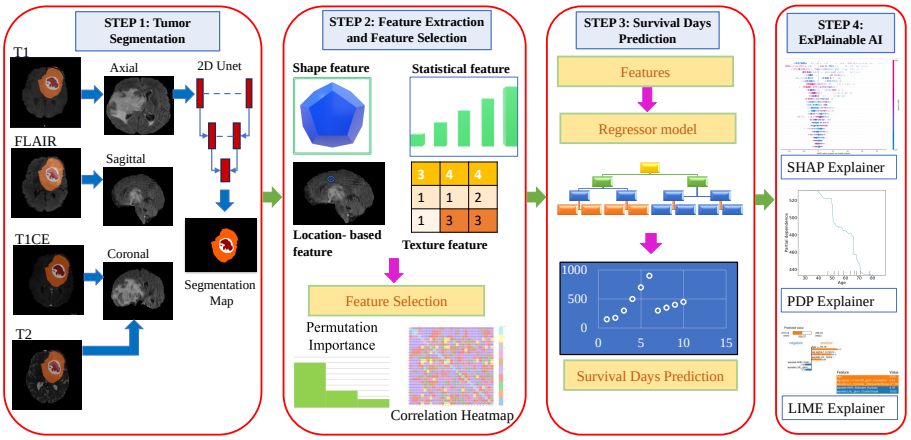


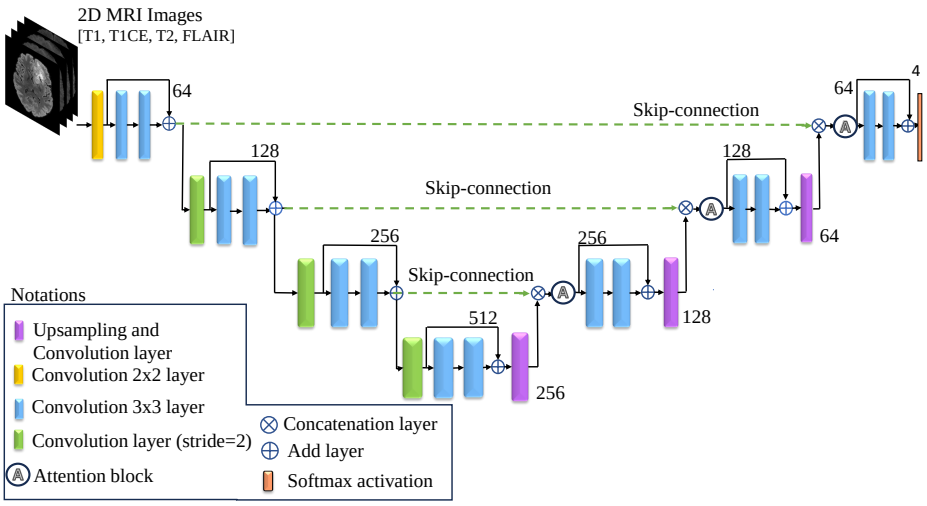
Fig. 1: The complete workflow of the proposed methodology for SD prediction.

3 Proposed Methodology for SD Prediction

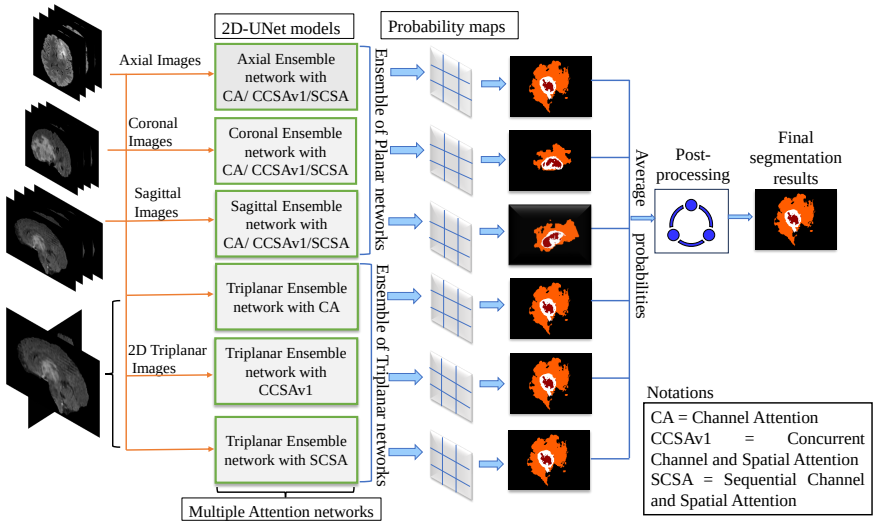
This study focuses on predicting SD in glioma patients, and the entire workflow of our proposed methodology is illustrated in Fig. 1. It emphasizes the critical role of BTS in predicting SD, depicted in Fig. 2 for the structural diagram. For our approach to SD prediction, segmentation results from the BTS task (discussed in Section 3.1) were used to extract features. In our study, we referenced [36] work where the authors emphasized the robustness of the SD prediction feature set. Consequently, we extracted the same features from the TN used in BTS. For further assessment of the robustness of these features, we employed various evaluation techniques, including correlation maps and post-hoc interpretability methods such as Shapley-additive explanations (SHAP) [22], accumulated local effect (ALE) [4], local interpretable model-agnostic explanations (LIME) [38], and partial dependency plots (PDP) [14]. This approach applies interpretation techniques to a pre-trained ML model to understand its decision-making process after training, enhancing transparency and explainability of the model's predictions.

3.1 Methodology for BTS

The work uses an ensemble of triplanar networks to generate segmentation results. It typically comprises three identical 2D UNet attention-based models [34, 35], each trained separately on specific axial, coronal, and sagittal image planes. The TN network ensemble is illustrated in Fig. 2b. The final segmentation outcomes are derived by integrating the outputs from these individual models illustrated in Fig. 3. Three unique lightweight attention mechanisms, namely channel based attention (CA), concurrent channel and spatial based attention (CCSAv1), and sequential channel and spatial based attention (SCSA), were integrated with the TN network to create its various variants [34, 35]. The basic architecture of 2D UNet is illustrated in Fig. 2a.



(a)



(b)

Fig. 2: (a) Basic 2D UNet architecture. (b) The structural diagram of the BTS network. Here, CA, CCSAv1, and SCSA are distinct attention mechanisms [34, 35].

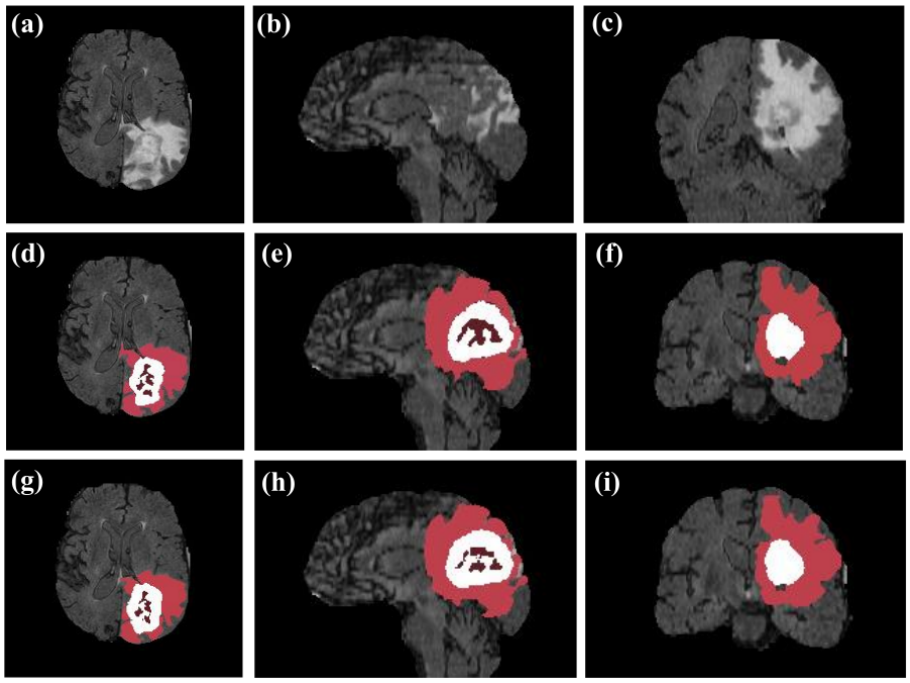


Fig. 3: (a), (b), and (c) display the axial, sagittal, and coronal planes of the input (FLAIR) image, respectively, while (d), (e), and (f) show the corresponding manual ground truth segmentation. Similarly, (g), (h), and (i) depicts the corresponding predicted segmentation maps for the planer view. The WT region is represented by light brown, white, and dark brown, the ET region by white, and the TC region by dark-brown and white [35].

Network implementation details The input MRI images are preprocessed by performing bias field correction using the N4ITK tool [46], removing non-brain pixels, eliminating the top and bottom 1% of intensity outliers, and normalizing the intensity of each image slice. The input image slice dimensions are 192×152 for the axial model, 192×152 for the sagittal, and 152×144 for the coronal model. In our training process, we applied random horizontal and vertical image flips as part of our data augmentation techniques.

Fig. 2a depicts the primary network employed for all the models. It has a basic encoder-decoder architecture. In the encoder stage, two convolution blocks are connected like a residual block connection. Here, downsampling is achieved through the strided convolutions. Whereas, the decoder stage includes an upsampling layer, a 2×2 convolution layer, a residual block, and an attention block. Lastly, a softmax activation produces feature maps for each class. The utilized loss function is a fusion of cross-entropy and generalized dice-loss functions. Segmentation results were postprocessed by applying connected components analysis to eliminate false positive regions from the segmented regions.

Evaluation metric Semantic segmentation networks are evaluated quantitatively using the Dice similarity coefficient (DSC) and 95% of Hausdorff distance (HD) metrics. The DSC measures the overlap of pixels between the predicted segmentation and the ground truth. Its value ranges from 0 to 1, with 1 representing perfect similarity and 0 indicating no overlap. HD is the greatest distance between a point in one of the two sets and its closest point in the other. The most significant segmentation error is indicated by HD, making it one of the most informative and helpful criteria [18]. The value falls within the range of 0 to 1, and a smaller Hausdorff95 signifies a higher segmentation quality, indicating that the predicted brain tumor boundary closely matches the actual boundary. DSC is defined as Equation 1 and HD is defined as Equation 2:

$$DSC = \frac{2TP}{FP + 2TP + FN} \tag{1}$$

$$\begin{aligned} hd(P,G) &= max_{p \in P}\ min_{g \in G}\ ||p - g||_2 \\ hd(G,P) &= max_{g \in G}\ min_{p \in P}\ ||p - g||_2 \\ HD(P,G) &= max(hd(P,G),\ hd(G,P)) \end{aligned} \tag{2}$$

Where FP, TP, and FN represent the counts of false positive, true positive, and false negative voxels, respectively. P represents the set of pixels in the predicted tumor, with p representing the pixel in set P. Similarly, G represents the set of pixels in the ground truth, with g representing the pixel in set G. $HD(P,G)$ is the hausdorff distance between the sets P and G. This measures the greatest distance between a pixel in one set and the nearest pixel in the other.

3.2 SD Prediction Methodology

Motivated by the work conducted by [36] on SD prediction, we generated 29 unique features derived from the TN segmentation network for training and validation samples. The authors extracted 1265 features, including statistical, shape, location, and texture information using wavelet and LoG filters, broadly categorized into images and radiomics features. Wavelet filters are widely used for image denoising. In contrast, Laplacian of Gaussian (LoG) filters, acting as generic differential operators sensitive to local image variations like edges or blobs, have enhanced the performance of SD prediction [9, 12]. Permutation importance (PI) [29] and SRC were utilized as feature selection techniques. PI is also an interpretation technique that evaluates the importance of all the features by measuring the increase in the model's prediction error when the value of the specific feature is permuted or shuffled. A feature is considered significant if altering its values increases the prediction error. It generates weights of all the features, where higher weights signify higher contribution in SD prediction, whereas 0 and negative define zero contribution. This provides insights into the feature's contribution to prediction and helps us understand its significance.

We extracted 29 features from the ground-truth (GT) images and an additional set from the predicted segmentation results from the TN network, and we used it for model training. We trained distinct Random Forest regressor (RFR) models using these feature sets. The parameters, including {*number of trees, minimum sample leaf, maximum depth, maximum features, and random state*}, were fine-tuned using the grid search technique. Following that, we explored the behavior of these features concerning the outcomes.

XAI - Post-hoc methods After training, these post-hoc methods interpret a model's decisions, explaining how input features influence output predictions in complex ML models. The SHAP summary plot can derive global and local explanations, whereas the force and waterfall plots derive local explanations from the model. These plots visually represent the contribution of each sample to the SD prediction. Aggregating the SHAP values allows one to find the contribution of each feature as it considers all possible combinations of features. Further, for each combination of features, SHAP determines a feature's contribution by measuring the change in the model's prediction when the feature is included versus when it is excluded, resulting in the SHAP value [36].

LIME can generate a local or individual interpretation of features; the core idea behind LIME is to create an approximate linear model centered on the explained example. This approximation is achieved by generating numerous synthetic examples near the explained instance, with each example weighted according to its distance from the explained instance. Using these generated examples, a linear regression model is built, with the coefficients serving as quantitative indicators of the influence of individual features on the prediction.

Conversely, PDP and ALE plots serve the purpose of extracting global insights from the model. PDPs calculate the average prediction of the model for a specific feature while keeping all other features fixed. They then vary the feature of interest across a range of values to observe how it impacts the model's predictions [30]. Hence, PDP provides a global perspective on how the selected feature influences the model's output across its entire range. However, there are two significant limitations of PDP:

- The assumption that the feature of interest is uncorrelated with other features.
- Accumulating marginal effects across all samples, neutralizing the heterogeneous effects of specific feature values.

These limitations are mitigated by ALE, which computes the average change in predictions as the feature of interest varies across its observed range. Rather than presuming all other features are constant, ALE considers the local effects within each segment and aggregates them across the entire feature range.

4 Results and Discussions

The DSC for segmentation results obtained from the TN network is 0.736 (ET), 0.841 (TC), 0.896 (WT) for the training set and 0.713 (ET), 0.778 (TC), 0.873

(WT) for the validation set. The results are consistent with many 2D and 3D UNet-based leading models [1, 3, 5].

In the context of SD prediction, the performance of the RFR models on the training and validation set can be seen in Table 2. The *RFR-GT* (RFR model trained on feature extracted from the ground-truth) model has performed better than the *RFR-TN* (RFR model trained on feature extracted from Triplanar network) model on both training and validation sets in terms of accuracy. The *RFR-TN* model shows lower errors in MSE, medianSE, and stdSE than the *RFR-GT* model on the training set. On the validation set, it also exhibits lower errors in MSE and stdSE. Regarding SRC, *RFR-TN* model has performed better on the training set, whereas *RFR-GT* performed better on the validation set. In a broader context, the *RFR-TN* model consistently performs better than *RFR-GT* across all the performance metrics.

Table 2: Performance evaluation on BraTS2020 training and validation datasets. Boldface numbers indicate the best outcomes. Where MSE= mean squared error, medianSE = median squared error, stdSE = standard deviation squared error, SRC = Spearman ranking coefficient, RFR GT= RFR model trained on GT feature set, and RFR TN = RFR model trained on TN feature set.

	RFR-GT					RFR-TN				
Dataset	Accuracy	MSE	medianSE	stdSE	SRC	Accuracy	MSE	medianSE	stdSE	SRC
Training	**0.590**	59961.29	14329.44	130263.47	0.75	0.540	**52490.06**	**13735.40**	**110568.72**	**0.84**
Validation	**0.607**	84583.28	**25863.77**	149488.19	**0.52**	**0.570**	**82070.60**	40678.11	**138345.10**	0.47

[a] Validation results from the BraTS2020 challenge online assessment portal: https://ipp.cbica.upenn.edu/

Since the *RFR-TN* model performs consistently on multiple metrics, we also want to assess the SD prediction for the TN segmentation model. Consequently, we generated a correlation map, visible in Fig. 4, to gain insights into the information carried by these features. These features exhibit little correlation, indicating that each feature holds distinct and valuable information. Within this map, correlations are confined to the range of -0.25 to +0.25, demonstrating that these features effectively capture distinct phenotype information from tumor lesions. The feature list of the correlation map can be found in Supplementary Table 1.A.2. Subsequent sections discuss further modeling of TN features using post-hoc interpretable techniques.

4.1 Insights into Local Interpretation and Biological Linkages

In alignment with our objective, we investigated the behavior of features from both a global and a local view (sample-wise) perspective using post-hoc interpretation tools. Fig. 5 illustrates a SHAP summary plot, where the X-axis represents the SHAP values, with their absolute values indicating their influence on the target feature (SD) and signs indicating their role in the increase and

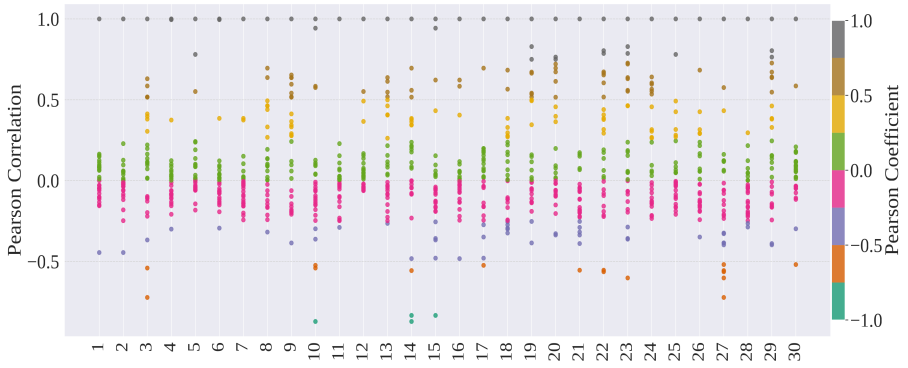


Fig. 4: Correlation map of feature set derived from TN segmentation network. (Refer to Supplementary Table 1.A.2 for features annotation.)

decrease of SD. On the Y-axis, features are arranged in increasing order of importance, whereas, on the right side of the plot, the color code illustrates the range of feature values, with high values depicted in pink and low values in blue. Each data point (sample), categorized by its feature value (either high or low), can be positioned along the X-axis to represent its impact on SD (whether it leads to an increase or decrease) and the magnitude of that impact.

Observing the summary plot, *Age* feature is identified as the most influential factor in predicting SD. When visualizing it on the SHAP value scale (X-axis), higher Age values (represented by green dots) correspond to a reduction in SD (indicated by negative SHAP values). In comparison, lower Age values (blue dots) contribute to increased SD (indicated by positive SHAP values). Additionally, the *Age* feature values exhibit an evenly distributed pattern. This observation is consistent with established medical knowledge emphasizing the crucial role of age in predicting survival duration for brain tumor patients [11, 20].

Similarly, the second feature, *LoG-sigma-1-0-mm-3D-Glcm-Correlation (Glcm _corr)*, is a texture feature that measures the linear association between the grayscale values of pixel pairs within an image. It plays a crucial role in image analysis, characterizing and differentiating various texture patterns found in images [48]. This feature is extensively used as a biomarker of heterogeneity allows it to offer insights into the tumor microenvironment, assisting in tumor classification [7, 32, 43]. Based on this plot, we can infer that a higher correlation is associated with an increase in SD, whereas lower values are linked to a reduction in SD. We argue that a low correlation indicates greater complexity (heterogeneity) among tumor microenvironments, decreasing SD. The feature, *cent_ncr_x*, represents the centroid of necrotic tumor lesions along the x-axis. A higher value of this feature is associated with a decrease in SD. It is a well-established fact that a tumor located in the posterior brain, particularly near the ventricle, has a detrimental impact on the patient's survival [23]. We can extend this observation to the remaining features by analyzing the interplay between

feature values, SHAP values, and the distribution of feature values. Similar to [36], we also found dominion of Age, location-based feature (*centroid of necrosis* and *active tumor*), *first-order kurtosis*, *glcm correlation*, and *gldm dependence variance*.

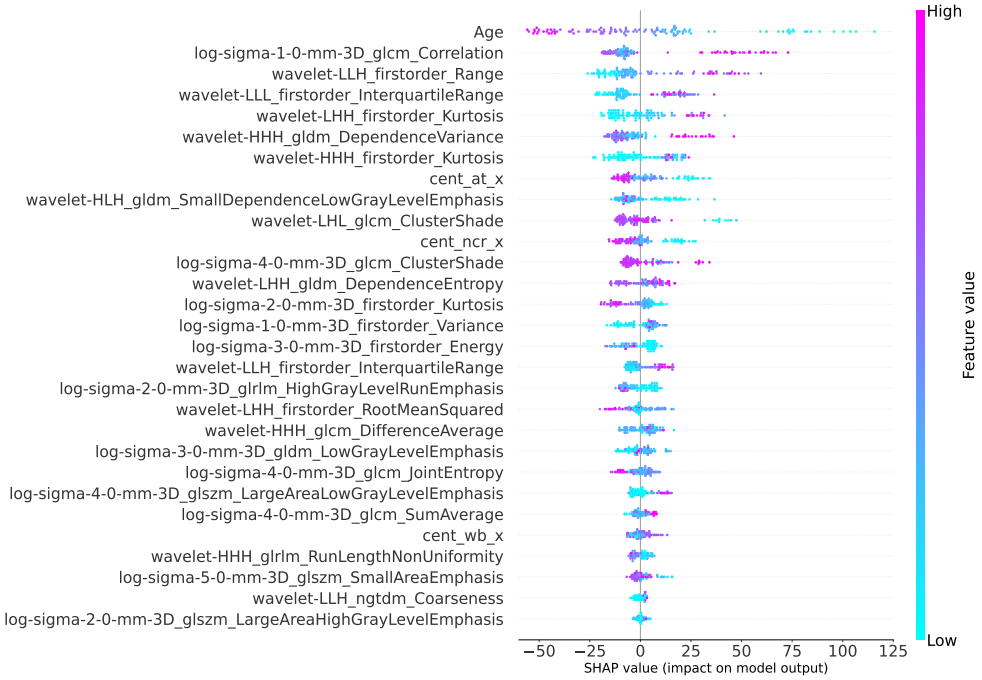
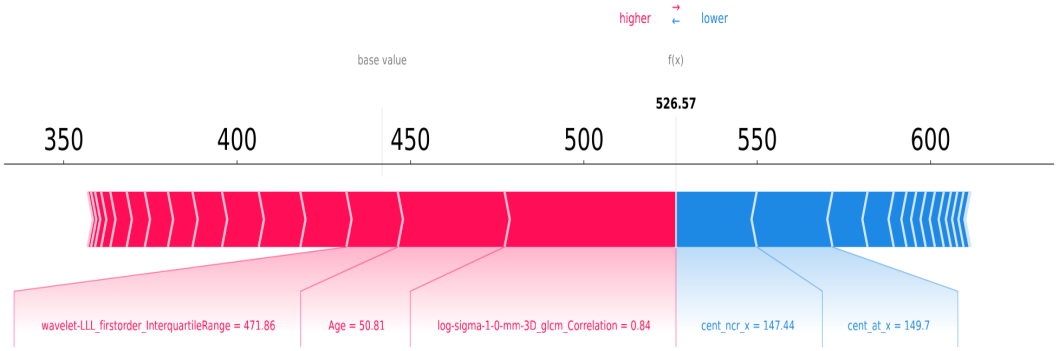


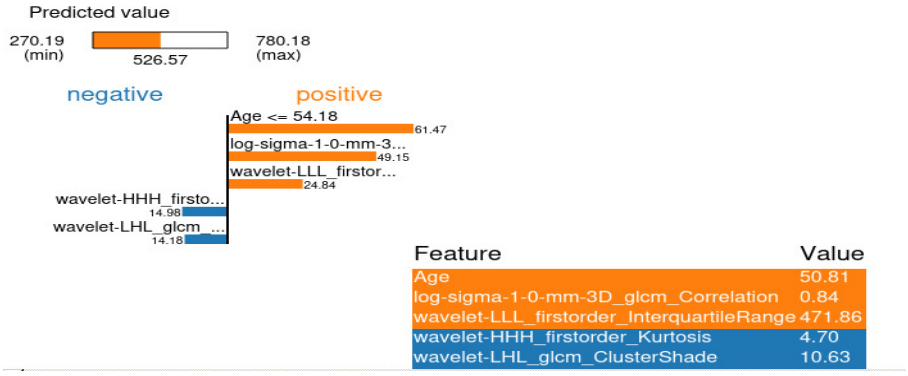
Fig. 5: The SHAP summary plots for the feature set display individual patients as blue dots. On the X-axis, the SHAP value unique to each patient can be observed, which signifies the influence of a particular feature on that patient's survival days. A larger absolute SHAP value indicates a more significant influence on survival days. In contrast, the sign of the SHAP value indicates whether it contributes to an increase or decrease in the average survival days. On the Y-axis, the features are arranged in descending order of importance. The color scheme on the right side of the plot is used to distinguish between high feature values depicted in pink and low feature values represented in blue.

Further, to explore the behavior at the sample (local) level, we employed the SHAP force-plot and LIME tools to visualize the feature behavior of a sample, which can be seen in Fig. 6a and Fig. 6b, respectively. The Python SHAP (version 0.42.1) and LIME tool (version 0.2.0.1) were employed for this purpose. In Fig. 6a, we displayed features whose contribution exceeds 8% (i.e., only features with a magnitude greater than 8% of the sum of all absolute Shapley values). Here, the size and color of the arrow signify the magnitude and direction (increasing/decreasing) of contribution to SD prediction. The red color represents an increase in SD, whereas the blue color indicates a decrease in SD. The LIME plot illustrated in Fig. 6b shows the top five contributing fea-

tures. Here, features and their respective value highlighted in orange indicate an increase in SD, while those in blue signify a decrease in SD. Comparing this Fig. 6b and Fig. 6b reveals that top-performing features i.e *Glcm_ corr*, *Age* and *wavelet-LLL_ firstorder_ InterquartileRange* each showing similar effects (increasing/positive) on SD prediction. However, the other two features responsible for reducing SD differ. Furthermore, we have included a SHAP waterfall plot for the same sample, available in Supplementary Figure 1, which shows the top ten features and indicates consistent behavior of features.



(a)



(b)

Fig. 6: (a) SHAP force plot for a sample illustrating features that contribute beyond an 8% threshold. Where *f(x)* represents the predicted value for this sample from the validation set of the training data. The baseline value is the expected value of the training set. Red-highlighted features indicate increasing SD, while blue-highlighted ones signify decreasing SD. The arrow size represents their predictive contribution. (b) LIME plot for a sample showcasing the influence of the top five features. Here, the predicted minimum and maximum values are shown, with orange indicating increased SD and blue indicating decreased SD. Features are arranged by decreasing importance.

4.2 Insights into Global Interpretation and Biological Linkages

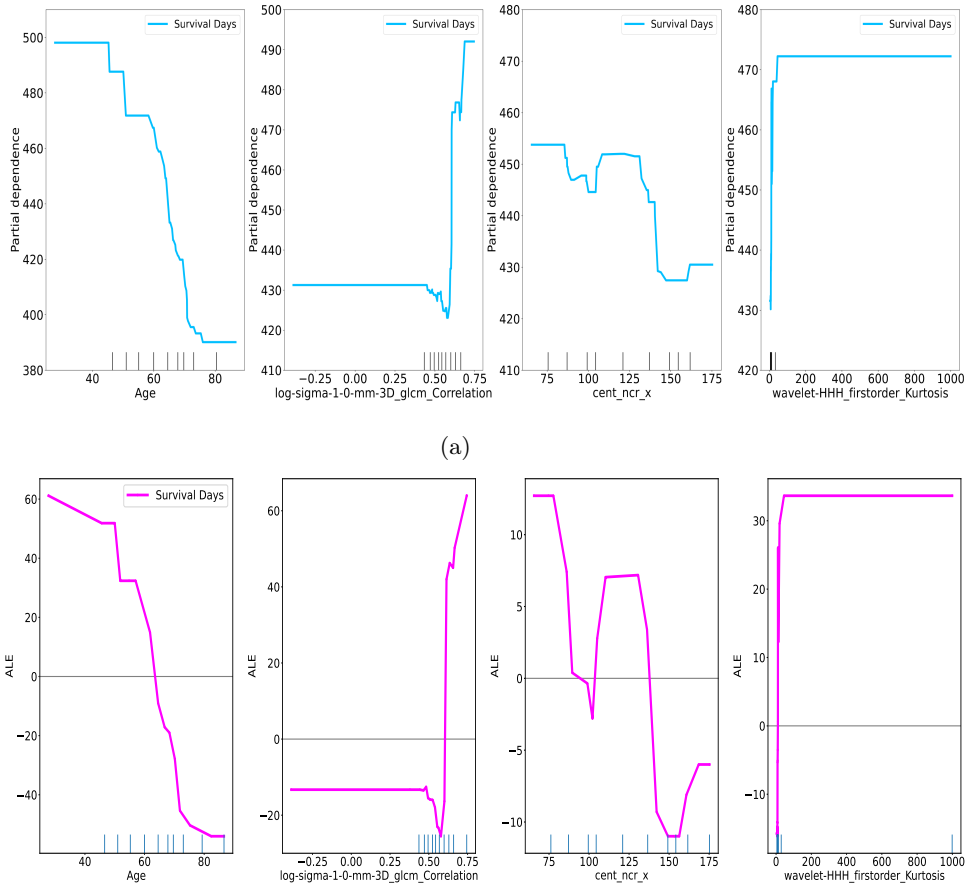


Fig. 7: Plotting first four features using (a) PDP and (b) ALE to analyze global behavior of the initial four features from SHAP summary plot. The X-axis represents feature values, whereas the vertical bar shows data distribution. The Y-axis in the PDP plots represents the average prediction changes in SD as the chosen feature changes. In the ALE plot, the Y-axis represents the accumulated change in the model's predictions as the desired feature changes while accounting for changes in the feature displayed on the X-axis within a set range.

We have employed Sklearn (version=1.1.1) [31] PartialDependenceDisplay tool to plot PDP and alibi (version=0.9.4) [19] ALE tool to analyze the global behavior of the initial four features from the SHAP summary plot (shown in Fig. 5). The PDP plot can be depicted in Fig. 7a, and the ALE plot in 7b. The X-axis in both the plot shows features and their respective values. In contrast,

the Y-axis in the PDP plot represents the average predicted SD as the chosen feature varies, while other features are held constant or averaged. In the ALE plot, the y-axis represents the accumulated local effect of the desirable feature on the predicted SD integrated over the specified feature range.

By observing the average changes in SD plotted on the Y-axis for PDP (Fig. 7a) and the accumulated changes from ALE (Fig. 7b) plots concerning the X-axis, we notice a consistent overall behavioral pattern for the desired features. For example, when visualizing the *Age* feature in a PDP plot, the most significant changes in the average effect, denoted on the Y-axis (SD:500), are observed within the 40-50 age range (shown on the X-axis). A greater magnitude of changes in the average effect signifies the feature's significance in predicting SD. This observation is similarly depicted in the ALE plot for the *Age* feature. Observing the vertical bar on the X-axis indicates the distribution of feature value. For *Age* feature, the distribution is even ranging from 40 to 80.

In the case of *Glcm_ corr*, most features exhibit correlation values ranging from 5 to 0.75, with one sample displaying a negative correlation. Similarly, for *cent_ ncr_ x*, most centroid coordinates of necrosis regions along the X-axis are between 7-170. Whereas for *wavelet-HHH_ firstorder_ Kurtosis*, most values ranging from 3 to 50 indicate high (positive) kurtosis, suggesting a concentration of the distribution toward the tails rather than the mean. In diffusion kurtosis imaging (DKI), kurtosis is a metric for evaluating tissue microstructure, offering insights into tissue barrier complexity and cellularity [40]. Deviations from typical kurtosis values may signal tissue integrity changes. Positive kurtosis correlates with increased tissue heterogeneity in ischemia and infarction [40]. A study found that high-grade tumors exhibited higher kurtosis values, likely due to greater cellular density, reduced cell size, and increased complexity (heterogeneity) in the tumor microenvironment [47].

Lastly, we have compared our best-performing model on all the performance metrics with the top-ranking and leading models in Table 3. The bold-faced text shows the best results. Our proposed RFR-TN model has surpassed the leading approach in terms of both accuracy and MSE. However, in terms of SRC, it ranks as the second-best performer.

5 Conclusions and Future Scope

We utilized the TN network to extract features for SD prediction. As claimed by the authors, the robustness and reliability of the feature set were achieved by extracting 29 dominant features from the ground-truth and TN-predicted segmented outcome to predict patient SD. We validate the effectiveness of these features by examining correlation maps and SRC from the two variants. Additionally, the *RFR-TN* model outperformed other top-performing BraTS2020 models across multiple performance metrics. Post-hoc interpretation methods have recently become critical and widely employed tools for explaining ML models. However, diverse post-hoc interpretation methods can yield different interpretations for SD prediction, raising questions about which method provides the

Table 3: Quantitative comparison of the proposed method's SD performance with leading models on the BraTS2020 training and validation datasets, utilizing data from the validation2020 leaderboard [39]. The bold-faced text shows the best results. The method highlighted in yellow is the top-ranking approach from the BraTS2020 challenge, while the method highlighted in grey reports results on the BraTS2020 validation dataset NA: Not Available.

Dataset	Method	Accuracy	MSE	medianSE	stdSE	SRC
Training	Mckinley et al. [25]	NA	NA	NA	NA	NA
	Asenjo et al. [5]	0.822	55499.71	11351.02	147319.00	0.833
	Bommineni et al. [8]	NA	NA	NA	NA	NA
	Ali et al. [3]	0.641	62305.61	05745.64	200788.00	0.632
	Rajput et al. [36]	0.538	60668.61	16037.10	125873.00	0.754
	RFR-TN(Proposed)	0.540	52490.06	13735.40	110568.72	0.84
Validation	Mckinley et al. [25]	0.414	098704.66	36100.00	152176.00	0.253
	Asenjo et al.[5]	0.520	122515.80	70305.26	157674.00	0.130
	Bommineni et al. [8]	0.379	093859.54	67348.26	**102092.00**	0.280
	Ali et al. [3]	0.483	105079.40	37004.93	146376.00	0.134
	Rajput et al. [36]	0.552	079826.24	**14148.89**	148288.00	**0.711**
	RFR-TN(Proposed)	**0.570**	082070.60	40678.11	138345.10	0.470

most accurate post-hoc interpretability. Therefore, we cross-evaluated the visual outcomes from various post-hoc interpretation methods for SD prediction. The visual derivations of features from these methods were consistent, indicating the robustness of the feature set.

We analyzed the global behavior of the feature set using SHAP, PDP, and ALE and the local behavior using SHAP and LIME. This approach allows us to comprehensively assess how each feature influences model predictions and provides insights into the collective impact and individual contributions of features, enhancing our understanding of model interpretability and robustness. Moreover, employing interpretability tools enables the extraction of human-understandable inferences, assisting in comprehending ML black-box models. For future work, we will test this feature set on the 3D segmentation model with explainability tools. Various studies have substantiated the importance of integrating clinical data such as age, gender, race, performance score, and treatment information to predict the SD of glioma patients.

Acknowledgments. M. Roy acknowledges the seed grant No. ORSP/R&D/PDPU/2019/MR/RO051 of PDEU (for the computing facility), the core research grant No. CRG/2020/000869 of the Science and Engineering Research Board (SERB), Govt. of India and the project grant no $GUJCOST/STI/2021-22/3873 of GUJCOST$, Govt. of Gujarat, India. M. S. Raval acknowledges the grant number. $GUJCOST/STI/2021-22/3858$ of Gujarat Council of Science and Technology (GUJCOST), Govt. of Gujarat, India for computing facility to support the work.

References

1. Agravat, R.R., Raval, M.S.: Brain tumor segmentation and survival prediction. In: International MICCAI Brainlesion Workshop. pp. 338–348. Springer, Shenzhen, China. (2019)
2. Akbar, A.S., Fatichah, C., Suciati, N.: Unet3d with multiple atrous convolutions attention block for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I. pp. 182–193. Springer, Strasbourg, France. (2022)
3. Ali, M.J., Akram, M.T., Saleem, H., Raza, B., Shahid, A.R.: Glioma segmentation using ensemble of 2d/3d u-nets and survival prediction using multiple features fusion. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6. pp. 189–199. Springer, Strasbourg, France (2021)
4. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society Series B: Statistical Methodology **82**(4), 1059–1086 (2020)
5. Asenjo, J.M., Solís, A.M.L.: MRI brain tumor segmentation using a 2d-3d u-net ensemble. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 354–366. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-72084-1_32
6. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
7. Biswas, R., Nath, A., Roy, S.: Mammogram classification using gray-level co-occurrence matrix for diagnosis of breast cancer. In: 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE). pp. 161–166. IEEE (2016)
8. Bommineni, V.L.: PieceNet: A redundant UNet ensemble. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 331–341. Springer International Publishing, Strasbourg, France (2021). https://doi.org/10.1007/978-3-030-72087-2_29
9. Chato, L., Chow, E., Latifi, S.: Wavelet transform to improve accuracy of a prediction model for overall survival time of brain tumor patients based on mri images. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI). pp. 441–442. IEEE, New York, USA (2018)
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer, Athens, Greece (2016)
11. Davis, F.G., McCarthy, B.J., Freels, S., Kupelian, V., Bondy, M.L.: The conditional probability of survival of patients with primary malignant brain tumors: surveillance, epidemiology, and end results (seer) data. Cancer: Interdisciplinary International Journal of the American Cancer Society **85**(2), 485–491 (1999)

12. Demircioğlu, A.: The effect of preprocessing filters on predictive performance in radiomics. European Radiology Experimental **6**(1), 40 (2022)
13. Feng, X., Tustison, N.J., Patel, S.H., Meyer, C.H.: Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. Frontiers in computational neuroscience **14**, 25 (2020)
14. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
15. Fusco, R., Granata, V., Grazzini, G., Pradella, S., Borgheresi, A., Bruno, A., Palumbo, P., Bruno, F., Grassi, R., Giovagnoni, A., et al.: Radiomics in medical imaging: Pitfalls and challenges in clinical management. Japanese Journal of Radiology **40**(9), 919–929 (2022)
16. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I. pp. 272–284. Springer, Strasbourg, France (2022)
17. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnu-net for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6. pp. 118–132. Springer, Lima, Peru (2021)
18. Karimi, D., Salcudean, S.E.: Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Transactions on medical imaging **39**(2), 499–513 (2019)
19. Klaise, J., Looveren, A.V., Vacanti, G., Coca, A.: Alibi explain: Algorithms for explaining machine learning models. Journal of Machine Learning Research **22**(181), 1–7 (2021), http://jmlr.org/papers/v22/21-0017.html
20. Ladomersky, E., Scholtens, D.M., Kocherginsky, M., Hibler, E.A., Bartom, E.T., Otto-Meyer, S., Zhai, L., Lauing, K.L., Choi, J., Sosman, J.A., et al.: The coincidence between increasing age, immunosuppression, and the incidence of patients with glioblastoma. Frontiers in pharmacology **10**, 200 (2019)
21. Liu, C., Ding, W., Li, L., Zhang, Z., Pei, C., Huang, L., Zhuang, X.: Brain tumor segmentation network using attention-based fusion and spatial relationship constraint. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6. pp. 219–229. Springer, Lima (2021)
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
23. Macyszyn, L., Akbari, H., Pisapia, J.M., Da, X., Attiah, M., Pigrish, V., Bi, Y., Pal, S., Davuluri, R.V., Roccograndi, L., et al.: Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro-oncology **18**(3), 417–425 (2015)
24. Marti Asenjo, J., Martinez-Larraz Solís, A.: Mri brain tumor segmentation using a 2d-3d u-net ensemble. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6. pp. 354–366. Springer, Lima, Peru (2021)

25. McKinley, R., Rebsamen, M., Daetwyler, K., Meier, R., Radojewski, P., Wiest, R.: Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. arXiv preprint arXiv:2012.06436 (2020)
26. McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Wiest, R.: Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6. pp. 401–411. Springer, Singapore, Singapore (2021)
27. Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H.: Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proceedings of the european conference on computer vision (ECCV). pp. 552–568. Springer, Munich, Germany (2018)
28. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
29. MIT, M.K., Lopuhin, K.: permutation_importance (Aug 1965), https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html
30. Molnar, C.: Interpretable machine learning. Lulu. com (2020)
31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
32. Pratiwi, M., Harefa, J., Nanda, S., et al.: Mammograms classification using gray-level co-occurrence matrix and radial basis function neural network. Procedia Computer Science **59**, 83–91 (2015)
33. Rajput, S., Agravat, R., Roy, M., Raval, M.S.: Glioblastoma multiforme patient survival prediction. In: Proceedings of 2021 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2021) Medical Imaging and Computer-Aided Diagnosis. pp. 47–58. Springer, Birmingham, UK (2022)
34. RAJPUT, S., KAPDI, R., RAVAL, M., ROY, M.: Multi-view brain tumor segmentation (mvbts): An ensemble of planar and triplanar attention unets. Turkish Journal of Electrical Engineering and Computer Sciences **31**(6), 908–927 (2023)
35. Rajput, S., Kapdi, R., Roy, M., Raval, M.S.: A triplanar ensemble model for brain tumor segmentation with volumetric multiparametric magnetic resonance images. Healthcare Analytics **5**, 100307 (2024)
36. Rajput, S., Kapdi, R.A., Raval, M.S., Roy, M.: Interpretable machine learning model to predict survival days of malignant brain tumor patients. Machine Learning: Science and Technology **4**(2), 025025 (2023)
37. Rajput, S., Raval, M.S.: A review on end-to-end methods for brain tumor segmentation and overall survival prediction. arXiv preprint arXiv:2006.01632 pp. 1–22 (2020)
38. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
39. Spyridon (Spyros), B.C.S.: Validation survival leaderboard 2020. https://www.cbica.upenn.edu/BraTS20//lboardValidationSurvival.html (2021), accessed: 2021-06-12
40. Steven, A.J., Zhuo, J., Melhem, E.R.: Diffusion kurtosis imaging: an emerging technique for evaluating the microstructural environment of the brain. American journal of roentgenology **202**(1), W26–W33 (2014)

41. Sundaresan, V., Griffanti, L., Jenkinson, M.: Brain tumour segmentation using a triplanar ensemble of u-nets on mr images. In: International MICCAI Brainlesion Workshop. pp. 340–353. Springer, Lima, Peru (2020)
42. Sundaresan, V., Zamboni, G., Rothwell, P.M., Jenkinson, M., Griffanti, L.: Triplanar ensemble u-net model for white matter hyperintensities segmentation on mr images. Medical image analysis **73**, 102184 (2021)
43. Tamal, M.: Grey level co-occurrence matrix (glcm) as a radiomics feature for artificial intelligence (ai) assisted positron emission tomography (pet) images analysis. In: IOP Conference Series: Materials Science and Engineering. vol. 646, p. 012047. IOP Publishing (2019)
44. Tarasiewicz, T., Kawulok, M., Nalepa, J.: Lightweight u-nets for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6. pp. 3–14. Springer, Lima (2021)
45. Turbé, H., Bjelogrlic, M., Lovis, C., Mengaldo, G.: Evaluation of post-hoc interpretability methods in time-series classification. Nature Machine Intelligence **5**(3), 250–260 (2023)
46. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: improved n3 bias correction. IEEE transactions on medical imaging **29**(6), 1310–1320 (2010)
47. Van Cauter, S., Veraart, J., Sijbers, J., Peeters, R.R., Himmelreich, U., De Keyzer, F., Van Gool, S.W., Van Calenbergh, F., De Vleeschouwer, S., Van Hecke, W., et al.: Gliomas: diffusion kurtosis mr imaging in grading. Radiology **263**(2), 492–501 (2012)
48. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer research **77**(21), e104–e107 (2017)
49. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 109–119. Springer, Strasbourg, France (2021)
50. Yip, S.S., Aerts, H.J.: Applications and limitations of radiomics. Physics in Medicine & Biology **61**(13), R150 (2016)
51. Zhang, S., Niu, Y.: Lcmunet: A lightweight network combining cnn and mlp for real-time medical image segmentation. Bioengineering **10**(6), 712 (2023)
52. Zhou, X., Li, X., Hu, K., Zhang, Y., Chen, Z., Gao, X.: Erv-net: An efficient 3d residual neural network for brain tumor segmentation. Expert Systems with Applications **170**, 114566 (2021)
53. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer, Granada, Spain (2018)

Appendix 1.A Supplementary:

1.A.1 Supplementary Figures:

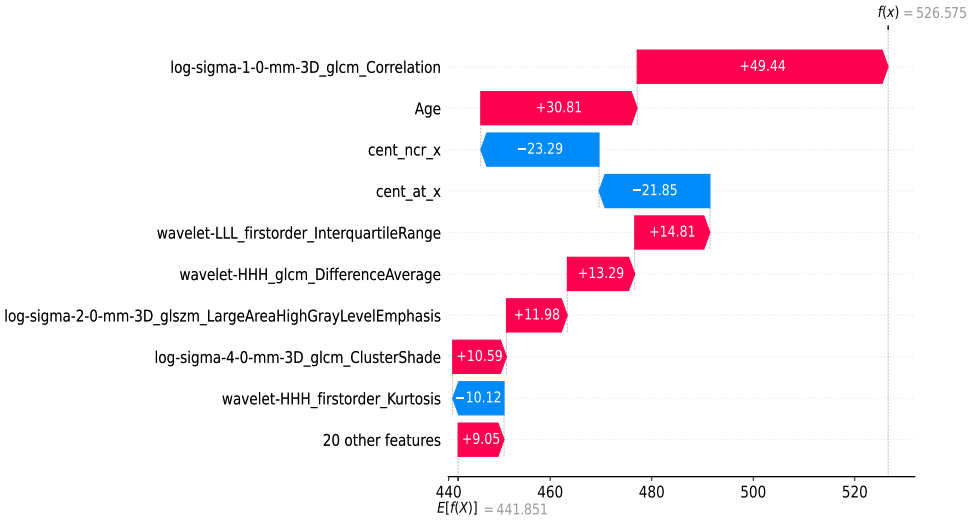


Fig. 1: The SHAP waterfall plot for the initial ten features illustrates the behavior of a sample from the validation set compared to the training set. Here, $E[f(X)]$ represents the expected (mean) value, $f(x)$ denotes the predicted value, and the direction of the arrow indicates whether the corresponding features contribute to increasing or decreasing survival days.

1.A.2 Supplementary Tables:

Table 4: Feature abbreviation from the correlation map.

Sr. No.	Feature Name	Feature Type	Sr. No.	Feature Name	Feature Type
1	*Survival Days*	Clinical Data (CD)	16	*Wavelet-HHH-Glcm-DifferenceAverage*	R
2	*Age*	CD	17	*Wavelet-HHH-Gldm-DependenceVariance*	
3	*LoG-sigma-1-0-mm-3D-Glcm-Correlation*	Radiomic based (R)	18	*Wavelet-HHH-Glrlm-RunLengthNonUniformity*	R
4	*cent-ncr-x*	Location-based (L)	19	*Wavelet-LLL-Firstorder-InterquartileRange*	R
5	*Wavelet-HHH-Firstorder-Kurtosis*	R	20	*LoG-sigma-1-0-mm-3D-FirstorderVariance*	R
6	*cent-at-x*	L	21	*LoG-sigma-4-0-mm-3D-Glcm-ClusterShade*	R
7	*cent-wb-x*	L	22	*LoG-sigma-4-0-mm-3D-Glcm-SumAverage*	R
8	*Wavelet-LLH-Firstorder-InterquartileRange*	R	23	*LoG-sigma-4-0-mm-3D-Glcm-JointEntropy*	R
9	*Wavelet-LLH-Firstorder-Range*	R	24	*LoG-sigma-3-0-mm-3D-Firstorder-Energy*	R
10	*Wavelet-LLH-Ngtdm-Coarseness*	R	25	*LoG-sigma-2-0-mm-3D-FirstorderKurtosis*	R
11	*Wavelet-LHL-Glcm-ClusterShade*	R	26	*LoG-sigma-2-0-mm-3D-Glszm-LargeAreaHighGrayLevelEmphasis*	R
12	*Wavelet-LHH-Firstorder-Kurtosis*	R	27	*LoG-sigma-3-0-mm-3D-Gldm-LowGrayLevelEmphasis*	R
13	*Wavelet-LHH-Firstorder-RootMeanSquared*	R	28	*LoG-sigma-4-0-mm-3D-Glszm-LargeAreaLowGrayLevelEmphasis*	R
14	*Wavelet-LHH-Gldm-DependenceEntropy*	R	29	*LoG-sigma-2-0-mm-3D-Glrlm-HighGrayLevelRunEmphasis*	R
15	*Wavelet-HLH-Gldm-SmallDependence-LowGrayLevelEmphasis*	R	30	*LoG-sigma-5-0-mm-3D-Glszm-SmallAreaEmphasis*	R