MACHINE
LEARNING
Science and Technology

PAPER

OPEN ACCESS

RECEIVED
19 January 2023

REVISED
15 April 2023

ACCEPTED FOR PUBLICATION
15 May 2023

PUBLISHED
30 May 2023

Interpretable machine learning model to predict survival days of malignant brain tumor patients

Snehal Rajput[1], **Rupal A Kapdi**[2], **Mehul S Raval**[3,*] and **Mohendra Roy**[1,*]

[1] SOT, Pandit Deendayal Energy University, PDEU Road, Gandhinagar 382007, Gujarat, India
[2] Institute of Technology, Nirma University, SG Highway, Ahmedabad 382481, Gujarat, India
[3] School of Engineering and Applied Science, Ahmedabad University, Commerce Six Roads, Ahmedabad 380009, Gujarat, India
[*] Authors to whom any correspondence should be addressed.

E-mail: mehul.raval@ahduni.edu.in and mohendra.roy@ieee.org

Keywords: explainable AI, machine learning, features importance, brain tumor segmentation, survival prediction, interpretability, Neural ODE

Abstract

An artificial intelligence (AI) model's performance is strongly influenced by the input features. Therefore, it is vital to find the optimal feature set. It is more crucial for the survival prediction of the glioblastoma multiforme (GBM) type of brain tumor. In this study, we identify the best feature set for predicting the survival days (SD) of GBM patients that outrank the current state-of-the-art methodologies. The proposed approach is an end-to-end AI model. This model first segments tumors from healthy brain parts in patients' MRI images, extracts features from the segmented results, performs feature selection, and makes predictions about patients' survival days (SD) based on selected features. The extracted features are primarily shape-based, location-based, and radiomics-based features. Additionally, patient metadata is also included as a feature. The selection methods include recursive feature elimination, permutation importance (PI), and finding the correlation between the features. Finally, we examined features' behavior at local (single sample) and global (all the samples) levels. In this study, we find that out of 1265 extracted features, only 29 dominant features play a crucial role in predicting patients' SD. Among these 29 features, one is metadata (age of patient), three are location-based, and the rest are radiomics features. Furthermore, we find explanations of these features using post-hoc interpretability methods to validate the model's robust prediction and understand its decision. Finally, we analyzed the behavioral impact of the top six features on survival prediction, and the findings drawn from the explanations were coherent with the medical domain. We find that after the age of 50 years, the likelihood of survival of a patient deteriorates, and survival after 80 years is scarce. Again, for location-based features, the SD is less if the tumor location is in the central or back part of the brain. All these trends derived from the developed AI model are in sync with medically proven facts. The results show an overall 33% improvement in the accuracy of SD prediction compared to the top-performing methods of the BraTS-2020 challenge.

1. Introduction

Brain cancer patients' survival rate is lower than other cancer types. The glioblastoma multiforme (GBM), or simply, glioblastoma, is the most invasive and frequently diagnosed type of brain tumor [1, 2]. Due to its infiltrative and diffuse characteristics, the World Health Organization (WHO) has categorized it as a Type-4 tumor [3]. Following the central-brain-tumor registry of the United States (CBTRUS)-2021 report, there were a total of 83 029 deaths in the USA alone between 2014 and 2018 due to malignant brain tumors and other central nervous system disorders tumors [2].

1.1. Brain tumor segmentation (BTS)

Usually, the brain anatomy analysis is done using magnetic resonance imaging (MRI) images, which are non-invasive, and provide high-resolution and detailed information about soft tissues. Recently, deep-learning-based approaches are becoming more popular for segmentation from medical images due to the introduction of powerful GPUs [4]. UNet-based approaches have generated robust segmentation results, as evidenced by their great performance in the medical image segmentation domain [5–7]. BTS separates cancerous tissues from healthy tissues, which can further dissect into necrosis, enhancing tumor (ET) or edema. In many standard benchmarks, such as in brain tumor segmentation challenge (BraTS) [8–10] counts the whole tumor (WT), tumor core (TC) and ET subregions for the evaluation of the segmentation methods.

The state-of-the-art BTS methods use 2D, 3D, or hybrid UNet [11]. The UNet performance is further improved by assembling attention blocks, [7, 12–15] residual connections between layers [16] and dense connections between layers of the network [6, 7, 17]. In the BraTS–2020, Isensee *et al* [18] proposed an improvised version of the 'No-New Network' model. Likewise, in the BraTS-2021 challenge, an optimized version of the same network was proposed at the conference of medical-image computing and computer-assisted intervention (MICCAI) 2021 [19]. The above segmentation techniques suggest that automatic segmentation is a complicated method due to the high variance in structure, shape, location, texture of tumor tissues, lack of ample images in the available standard dataset, and an imbalance between cancerous and healthy tissues. Thus, a robust segmentation method is desirable to develop an accurate and transparent survival prediction system.

1.2. SD prediction

The SD prediction is far more complex as it depends on many factors such as accurately segmented brain tumor [20], ample dataset, clinical information such as age, gender, health condition, treatment, biological characteristics, and qualitative image properties from radiographic images [21]. Though hugely challenging, it is crucial to improve early diagnosis, treatment planning, and post-treatment analysis of GBM patients [22, 23]. The GBM patients have a dismal survival record, with a median chance of survival of fewer than 12 months [24]. Various studies also show that the survival of patients varies with their age [2, 20, 25]. SD prediction from the BraTS challenge can be further categorized into long-term survival (where SD are >450 days), mid-term survival (300 to 450 days), and short-term survival (<300 days). Here, accuracy and Spearman ranking coefficient (SpearmanR) are used to evaluate the performance of the models.

1.3. End-to-end methods for BTS and SD

Since both the tumor segmentation and SD prediction are individually complex, therefore, various research groups are trying to develop an end-to-end model by integrating a tumor segmentation with the SD prediction method to make the system smooth and less complicated for SD prediction [26]. In this regard, Mckinley *et al* [7] proposed a 3D-2D densely connected encoder-decoder architecture for the segmentation task and thereby extracted the features. An ensemble of linear regressor and random forest classifiers was trained using age and features extracted from BTS to predict SD. Bommineni [27] proposed four identical networks for segmentation, where networks were trained on each class label and multiple class labels. They used the linear regressor for SD prediction and trained the model on the surface area, volume, spatial location, age, and resection status features.

1.4. Interpretability

Usually, BTS and SD models are not tested for their interpretability. In this regard, an end-to-end model that combines automated segmentation, feature extraction, and survival prediction with interpretability is a promising option. For the BTS task, we implemented 3D U-Net [28]. The features were extracted from segmentation results using various wavelet-based, location-based, shape-based, and radiomic-based filters. The radiomic features provide valuable insights into GBM prognosis but will be limited in providing biological insights. The several reasons for these limitations include—tumor heterogeneity, imaging limitation, and, most importantly, the lack of biological context. They can provide insights into the phenotypic structure but cannot explain the underlying molecular processes. Integrating radiomics with genomics, proteomics, or clinical data is necessary for a holistic view. This task is very complex and requires heavy computational resources and expertise. Therefore, the present work examines interpretability from the phenotypic perspective based on publicly available BraTS 2020 challenge data [29].

In addition, we used recursive feature elimination (RFE), PI, and correlation matrix to reduce the number of features. Further, we studied the correlation map, partial dependency plots (PDP) [30, 31], shapley additive explanations (SHAP) plots [32, 33], and Kaplan–Meier (KM) plots [34] to analyze the predictions. SHAP identifies the most important feature contributing to the prediction. This can aid the clinician in understanding the decision-making process and making treatment-related decisions. PDP will help to visualize how a particular radiomic feature affects prediction across different patients. This establishes the relationship between radiomic features and prediction and also reveals nonlinear dependencies amongst features. Thus, both can help make informed decisions and offer valuable insights into GBM prognosis.

In summary, our work focuses on the points listed follows:

- Finding an optimal feature set that augurs well for SD prediction.
- Validation of SD prediction on the BraTS-2020 dataset.
- Providing detailed explanations and rationale for the selection of the dominant features set.
- Interpretation of the model behavior and biomedical inference of the top six most important features.

All the obtained results are validated through the BraTS-Challenge-2020 evaluation platform [29].

2. Methods

2.1. End-to-end approach for SD prediction

The structural diagram of the proposed end-to-end approach is shown in figure 1. The multiple parametric MRI images are the input to the model such as T1-weighted (including contrast agent), T2-weighted, and fluid-attenuated inversion recovery (T2-FLAIR) images. The segmentation model is built on 3D U-Net architecture, known as the 'No-new Network' [28]. The architecture relies on 3D UNet, which is a well-proven architecture for biomedical segmentation tasks and is robust for tumor segmentation. The network consists of a symmetric five-layered encoder and decoder structure. It is a simple, easy-to-implement architecture with 8.3 million parameters. This makes it suitable for the resource-constrained 16 GB GPU and 256 GB RAM environment while maintaining good segmentation performance on BraTS 2020 dataset. For detailed architecture, please refer to supplementary figure A1. For this segmentation model's training, patches with sizes of 128 × 128 × 128 are randomly selected from the training dataset. The obtained mean Dice scores for Region of interest (ROIs) are 0.819 (WT), 0.766 (TC), 0.702 (ET) for BraTS2020 training set and 0.880 (WT), 0.858 (TC), 0.759 (ET) for validation set respectively.

The network and segmentation of the tumorous tissue from the training set are shown in figure A1(a) of the supplementary section. In addition, figures A1(b)–(d) also exhibits a qualitative comparison between the given input (T2-FLAIR) MRI image predicted image and ground truth. The SD predictor model was trained using the dataset's segmented results and ground truth. In contrast, it was tested on the features extracted from the segmented results of the validation set. The feature selection module finds the best group from these extracted features, which are then used to predict SD. Finally, the SD prediction module is investigated for its decision, understanding its generic (global/overall) and specific (local/sample-wise) behavior on SD. The details of the feature extraction, a feature selection module, the survival prediction model, and its interpretability are discussed in the subsections below.

2.2. Feature extraction module

The feature extraction module obtained the image-based features [25] and radiomics-based features [35] (table 1 lists the specifics of the features).

Here, the image-based features are extracted by determining the tumor's shape and location. In contrast, radiomics-based features are extracted from necrotic and non-enhanced tumor regions using wavelet and Laplacian of Gaussian (LoG) filters (with σ value 1 to 5). Here, the lower value of σ highlights fine textures, and the higher σ focuses on coarse textures. The wavelet filters denoise the images and capture spatial and global signals [36]. The LoG filter pinpoints the blob centers and approximates its size, shape, and orientation [37]. Thus, we obtained 1264 features (1225 radiomics-based + 39 image-based). We also considered the metadata, e.g. the age of patients, as a feature. As a result, 1265 features in total are being taken into account for the evaluation. Since some of these features can be redundant or not contribute to the prediction, a feature selection procedure is essential.

2.3. Feature selection module

The primary goal of feature selection methods is to eliminate unimportant or repetitive features. Here, we employed RFE [38] and PI [39] as feature selection methods. RFE is a backward feature selection method that re-fits the model after iteratively ranking the features according to their importance and eliminating the least important features. The description of the chosen dominant features identified by RFE are shown in supplementary table A1. On the other hand, PI finds influence in the model score by randomly re-arranging a single feature value. The pseudo-code of PI is shown in algorithm1. This technique breaks the connection between the desirable feature and the output feature. The model's score decline demonstrates how largely it depends on that feature. Thus, we weighted the features according to their importance. In general,

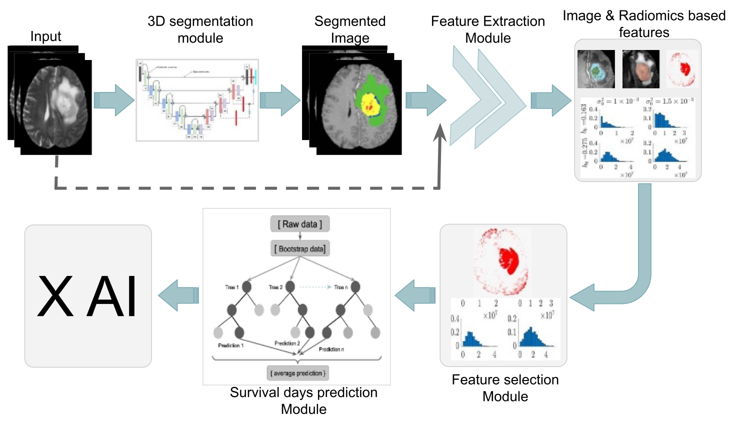


Figure 1. The workflow of the proposed end-to-end approach for the SD prediction.

Table 1. Feature-set lists 1264 features (1225 radiomics based + 39 image based).

	Image-based features
Shape-based features (27)	Surface area of ROIs, the volume of ROIs, proportion of ROIs, proportion ratio between each ROI, the area-to-volume ratio of ROIs, and amount of tumor.
Location-based features (12)	Centroid of ROIs, the distance between the center of ROIs and the center of the brain.
	Radiomics-based features
Shape features (13)	Elongation, major axis length, least axis length, mesh volume, flatness, maximum diameter row, maximum diameter column, surface area, sphericity, and surface volume ratio.
First-order statistical features (144)	Energy, maximum intensity value, minimum intensity value, mean, entropy, absolute deviation, inter-quartile range, variance, skewness, percentile, kurtosis, uniformity, and median.
Gray-level features (1068)	Neighboring gray-tone difference matrix (NGTDM), gray-level co-occurrence_matrix (GLCM), gray-level_size-zone (GLSZ), gray-level run-length matrix (GLRLM), and gray-level_dependence_matrix (GLDM).

Note: The values within the parenthesis represent the number of features extracted.

dominating features are given greater weight than other features. Zero or negative weights indicate no contribution of the feature for the prediction. Therefore, we removed them, bringing the set down to 180 features. These 180 features were further examined using the Spearman correlation coefficient (SpearmanR) with an absolute cutoff value of 0.5. It is clear from the correlation values that the necrotic, active, and WT centroids are firmly connected, given that they have similar characteristics in common. As a result, we narrowed the set of features to 29 by eliminating redundant features.

The pipeline for feature selection is as follows:

- We eliminated the features based on the PI weights (which define their contribution to the outcome). The threshold value of the weights is 100. Any features with PI weight <100 are eliminated. This results in 180 prominent features.
- Further, we eliminate the weaker features from these 180 features by finding the SpearmanR and a sorting process. For this, (a) we take features one by one (from the 180 feature set), starting with the feature having the least PI weight, and find its SpearmanR with the rest of the 179 features, (b) then we select the features which are having correlations less than 0.5, (c) then from this selected features, we identify the feature which is having highest PI weight value and use it to replace the feature that is having least PI weight (that we chose in step (a)). This process is repeated for each feature in the 180 feature set. That means the loop will run 179 times. Lastly, we find 29 dominant features (having less correlation) out of 180 features.

A detailed description of the selected dominant features using PI is shown in supplementary table A2.

2.4. Survival prediction module

The random forest regressor (RFR) [40] is based on ensemble learning, where decision trees (DT) are fundamental building blocks. Each DT was created using random samples from the training set; hence it is called a random forest. This method is widely used as it has been proven accurate and robust [40] across multiple complex problems, including SD prediction [41, 42]. The RFR model is often more successful than other models because the outcomes obtained by averaging the prediction from each tree result in lower variability. Additionally, randomization during tree growth and splitting helps prevent overfitting [43]. Hence, the RFR model is robust for predicting brain tumor patients' survival [44]. Here, a five-fold cross-validation technique was used to train the RFR model. Also, the hyper-parameters of the model were fine-tuned using grid search. The fine-tuned parameters are the *maximum tree depth, maximum number of features at each split, number of trees, and the minimal sample size required to be at a child node at a split point.*

Algorithm 1. Permutation importance (PI) algorithm.

```
Input: Trained model $m$ on the Dataset $D$
Compute: The metric $S$ of the model $m$ on dataset $D$ (for instance $R^2$ metric for a regressor model)
for each feature $j$ : do
    for each repetition $k$ in $K$: do
        Arbitrarily re-arrange column $j$ of Dataset $D$ to produce a noisy variant of the dataset say, $\hat{\mathrm{D}}_{k,j}$.
        Measure the metric $s_{k,j}$ of model $m$ on variant Dataset $\hat{\mathrm{D}}_{k,j}$.

    end for
    Measure importance $I$ of each feature $j$ defined as:
    $I_j = s - \frac{1}{k}\sum_{k=1}^{k} s_{k,j}$
end for
```

2.5. Interpretability methods for the proposed SD module

Understanding the decisions taken by AI or machine learning (ML) models is essential. Especially in the medical domain, the interpretability of such an AI model is vital to increase its reliability. Generally, the non-linearity in an AI model makes them hard to decipher. That is why we use model-agnostic methods like SHAP [32, 33] and PDP [30, 31] to find the interpretability of the proposed model.

The primary objective of the SHAP method is to determine how much each feature impacts the prediction for a given instance. The SHAP-value of a feature is the average marginal contribution (MC) of that feature to the value of the predecessor set among all possible permutations of the feature set. It can be expressed as in equation (1) [45].

$$(\Phi_j) = \frac{1}{\mid \Pi(N) \mid} \sum_{\pi \;\epsilon\; \Pi(N)} \overbrace{(v(\hat{P}_j^{\pi} \cup j) - (v(\hat{P}_j))}^{\substack{\text{marginal contribution of feature j} \\ \textit{in a coalition } \pi}} \tag{1}$$

where, (Φ_j) is the SHAP-value of feature of interest j, $\Pi(N)$ is the possible coalitions of all feature sets, π is the specific coalition, feature of interest is j, v is contribution of feature(s), $(\hat{P}_j^{\pi} \cup j)$ is the predecessor set of feature j in a particular coalition, including the j feature whereas $\hat{P}_j$ is predecessor set of feature j in a particular coalition, excluding j feature. E.g. if $\pi = \{A, B, D\}$, $j = B$ and $v\{A\} = 8$, $v\{B\} = 10$, $v\{C\} = 9$, $v\{A, B\} = 18$, $v\{A, D\} = 20$, $v\{B, D\} = 22$ and $v\{A, B, D\} = 25$, whereas the possible predecessor sets in this example) in a particular coalition $\pi = \{A, B, D\}$:$\{\phi, A\}$ and MC of $j(= \mathrm{B})$ is calculated as: $v\{A, B\} - v\{A\} = 20 - 8 = 12$. Further, calculating the MC of feature j across all the possible coalitions and averaging will give us a SHAP value (Φ_B) of feature B. In summary, it shows each feature's influence on predicting SD. It helps to understand the global behavior of the model by combining the explanation of each sample (please see the supplementary table A4 for a more detailed explanation of this example). Algorithm 2 displays the pseudo code to find the SHAP value for a feature.

The PDP displays the global effect of the feature on the target. The PDP considers all the samples and can show and examine the global association between SD and input variables. The partial dependence function is represented as :

$$f(x_s) = E_c[f(x_s,\, x_c)] \tag{2}$$

where x_s are the desirable feature(s) for which we want to plot partial dependency (PD) and x_c are the remaining features used to train the model. $x_c = x_s'$ and $X = x_s + x_c$ is the whole feature set. In PDP plot, we assume that feature subset x_s and x_c are uncorrelated to each other and hence can be calculated using average interaction effect [31] as:

$$f(x_s) = \frac{1}{n}\sum_{i=1}^{n} f(x_s,\, x_c) \tag{3}$$

Algorithm 3 displays the pseudo code to find the samples' PD values.

Algorithm 2. Calculating SHAP-value for a feature.

```
Input: Number of feature $N$ and their respective real value $v$ signifying their contribution. The contribution vector $v$ of a particular feature is calculated through perturbation feature values of coalition $\pi$. More details can be found here [46]. $k$ is the number of sampling permutations

Output: SHAP value $\phi_j$ for the feature $j \epsilon N$.
for Iteration : 1, 2, … K: do
    Randomly select $\pi$ from set of all permutation $\Pi(N)$
  for $j \epsilon N$ : do
      Calculating predecessor set $P_i^{\pi} = \{j \epsilon N \mid \pi(j) < \pi(i)\}$.
      $\phi_j = \phi_j + \frac{v(\hat{P}_j^{\pi} \cup j) - (v(\hat{P}_j))}{K}$
  end for
end for
```

Algorithm 3. The steps of obtaining PD value of samples are.

```
Input: The unique feature's values $x_s = x_1, x_2, \ldots x_n$, where $x$ is feature of interest

Ouput: PDP of desirable feature.
 Steps:
for i ϵ (1, 2, … , k): do
    Replace the original $x_1$ values with the constant $x_{1i}$ in the training samples.
    computes the predicted value vector from the altered copy of the training samples.
    compute the average of the prediction to find $f'(x_{1i})$.
end for
The PDP for $x_1$ is obtained by plotting the pairs $\{x_{1i}, f'(x_{1i})\}\ for\ i = 1, 2, ..n$
```

2.6. Performance metrics

Using multiple metrics for the performance evaluation provides the robustness information of the employed model. Hence, we quantified our model predictions on widely used metrics for survival prediction, such as accuracy [28, 42], mean squared_error (MSE) [28, 42], median squared_error (medianSE) [28, 42], standard-deviation standard_error (stdSE) [28, 42, 47], Spearman ranking coefficient [28, 42, 48].

2.7. Dataset BraTS-2020

The training BraTS 2020 [8–10] dataset includes 369 3D MRI samples for the segmentation and metadata (resection status information, Age, and SD). Out of this, 236 patients' metadata are provided for the SD prediction task. The validation BraTS 2020 dataset contains 125 MRI sample images and metadata of 29 patients. Each sample instance includes the fluid-attenuated-inversion recovery (T2-FLAIR), T2-weighted MRI preoperative images, T1 weighted (T1), post-contrast T1-weighted (T1-ce), and corresponding ground truth. In addition, the dataset is skull-stripped, aligned to the identical anatomical structure, and re-sampled to an isotropic resolution. The segmentation class labels, as defined in the BraTS-2020, are label-0 for background voxels, label-1 for necrotic and non-enhanced tumor voxels-(NCR or NET), label-2 for edema voxels-ED, and label-4 for ET voxels-ET.

3. Results and discussion

In this work, the prediction model was evaluated through the BraTS evaluation platform [29]. In addition, we have used the BraTS-2020 top-performing models as benchmarks to compare our results. Finally, this section discusses the results of the proposed end-to-end model for its performance and interpretability.

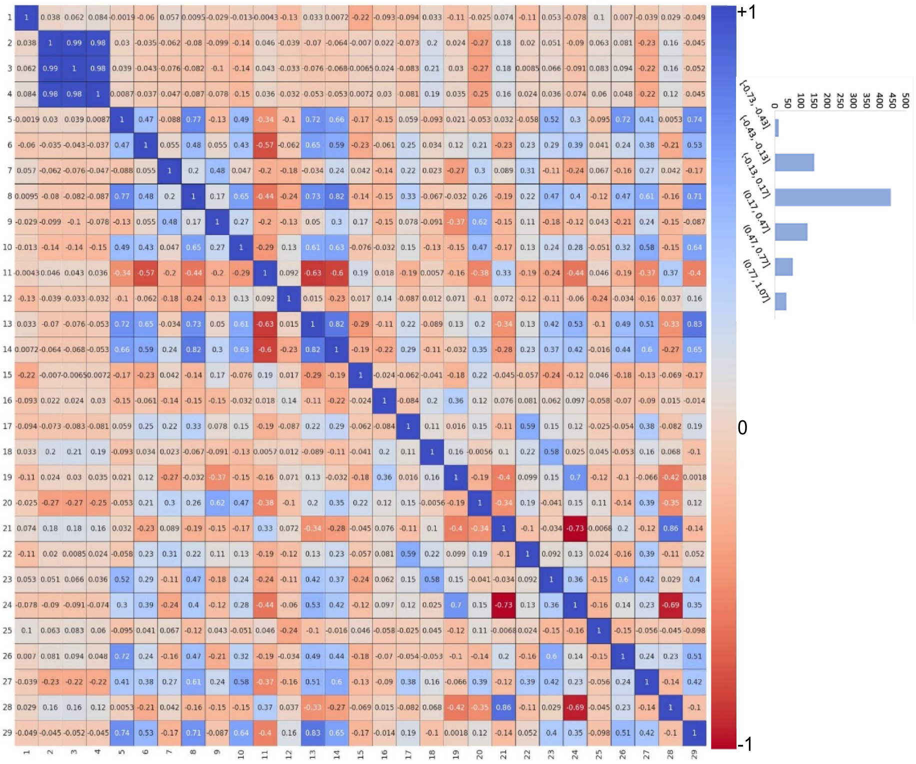


Figure 2. Correlation matrix of feature-set obtained through PI method. The histogram plot on the right-hand side depicts the range and the count for all the correlations in the heat map. (Refer supplementary table A3 for features annotation).

3.1. Correlation study of dominant features

To gain a better insight, we have plotted the correlation matrix of the features as shown in figure 2 (refer to supplementary table A3 for the annotation of the features). The plot shows that most features are highly uncorrelated, which signifies that they have captured distinct properties of phenotypes. Furthermore, the histogram in figure 2 validates that most of the selected features correlate from -0.13 to $+0.17$, suggesting they are uncorrelated, and it justifies the merits of our chosen features.

3.2. SD prediction results

The comparison of our SD prediction results with top-ranking methods of BraTS 2020 are shown in table 2. A robust method must perform well on multiple performance metrics apart from accuracy, as each quantifies the models on different parameters. Hence, we compared the proposed model with benchmark models [7] and reported the improvement as computed using equation (4). Here the percentage of improvement ϕ for each performance metrics x for our proposed model P given by:

$$\phi(x) = \frac{\text{Proposed\_model}(P) - \text{Top\_ranking\_model}(S)}{\text{Top\_ranking\_model}(S)} \times 100. \tag{4}$$

With this, the survival prediction result of the proposed method shows a 33.33% improvement in accuracy. There is a 19.13% improvement in MSE, which measures the variance around the fitted regression and indicates the deviation of model prediction from the actual one. However, it is sensitive to outliers [49]. In the case of median SE, there is a 60.80% improvement, which uses the median value of the residuals and is unaffected by the outliers. All these results obtained using various metrics indicate the robustness of the prediction [49]. We can see a 2.62% improvement in stdSE and a 181.03% improvement in SpearmanR coefficient often used to measure the relation between the therapy response and the SD [48]. As shown in table 2, our model has performed consistently in all the standard metrics used for SD prediction.

3.3. Interpretability of SD prediction model

This section presents a detailed analysis of the influence of features on SD prediction. The SHAP results provide local and global impact details, whereas PDP helps analyze features' global impact.

Table 2. SD performance comparisons with top-ranking models on the training and validation datasets BraTS 2020. The numbers of other models are obtained from the validation leader-board [29]. NA: not available.

Dataset	Method	Accuracy	MSE	medianSE	stdSE	SpearmanR
Training	Mckinley *et al* [7]	NA	NA	NA	NA	NA
	Asenjo and Solís *et al* [15]	0.822	55 499.71	11 351.02	147 319.00	0.833
	Bommineni *et al* [27]	NA	NA	NA	NA	NA
	Ali *et al* [50]	0.641	62 305.61	05 745.64	200 788.00	0.632
	Proposed method	0.538	60 668.61	16 037.10	125 873.00	0.754
Validation	Mckinley *et al* [7]	0.414	098 704.66	36 100.00	152 176.00	0.253
	Asenjo and Solís [15]	0.520	122 515.80	70 305.26	157 674.00	0.130
	Bommineni [27]	0.379	093 859.54	67 348.26	**102 092.00**	0.280
	Ali *et al* [50]	0.483	105 079.40	37 004.93	146 376.00	0.134
	Proposed method	**0.552**	**79 826.24**	**14 148.89**	148 288.00	**0.711**

Note: The bold text signifies the highest value in the respective performance metrics.

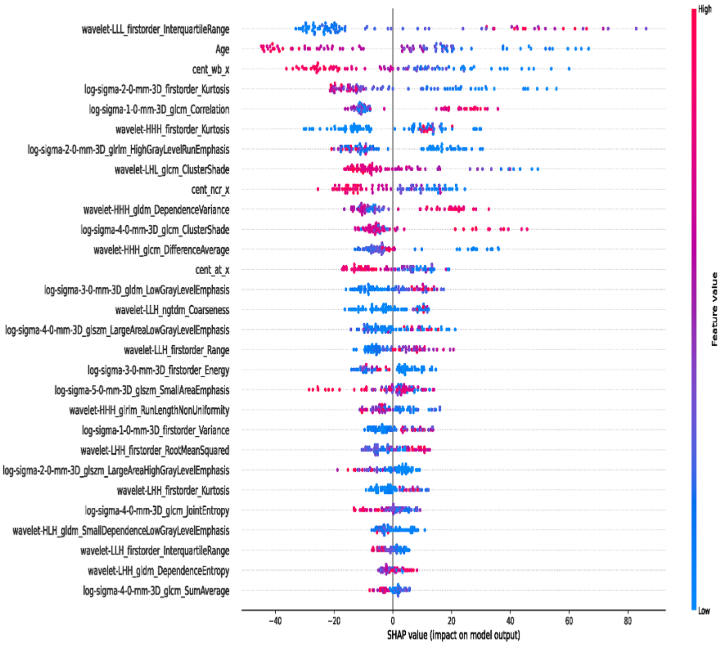


Figure 3. SHAP summary plot of dominant features: each blue dot represents a single patient. The X-axis shows the shapley (SHAP) value of a specific patient, which signifies the feature's effect on the survival days for the particular patient. The absolute higher SHAP value indicates a higher impact on survival days, where a sign indicates increasing or decreasing average survival days. The Y-axis shows features arranged according to importance (high to low) calculated through the average of absolute SHAP-value. On the opposite side, feature value shows high (red) or low (blue) values.

3.3.1. SHAP analysis results

SHAP depicts the importance of features in predicting a sample by calculating SHAP values. It shows the contribution of features to the expected prediction among all the feature combinations. The SHAP value shows how much a single feature affects the forecast, whereas the signs indicate whether the impact is positive or negative on the prediction outcome. Figures 3 and 5 show the SHAP summary and waterfall plots, respectively. The SHAP-summary plot helps us to visualize the global (generalize) and local (as it plots for every sample) impact of features on the model. In contrast, the waterfall plot allows us to visualize and study the features' impact on an individual sample. It will enable us to explore the role of features and their value on the particular prediction, where we can minutely examine each feature behavior for any desirable sample. In the SHAP-summary plot, X-axis displays the SHAP value, which signifies the impact of features on the target feature (here, the target feature is the *SD*). The greater the value (absolute), the more significant the effect on the target component, whereas the sign $(+/-)$ indicates whether that impact is positive or negative. In the Y-axis, features are listed in the order of importance (from top to bottom). Each point on the summary plot represents a sample, and the point's color represents the value of the corresponding instance. Here, blue denotes a low feature value, and red a high one.

From figure 3, we observe that *Wavelet-LLL_firstorder_InterquartileRange* (WIR) feature has the highest importance. It is a first-order radiomic feature extracted using the wavelet low pass filter and depicts the distribution of specific pixel values. *WIR* measures the pixel intensity between the 25% to 75% percentile range. From the plot, we can observe that the samples with intermediate or high feature values (purple and red color) of *WIR* contribute positively to the prediction, which has a maximum positive SHAP value. In other words, the intermediate or high feature value of the*WIR* feature increases the SD of patients. It is also apparent that there is an aggregation of large samples (with blue color) within the SHAP value range of -15 to -25 (refer to the *WIR* feature row listed on the Y-axis). It signifies that the majority of the samples fall into this SHAP range. The samples within this range are responsible for reducing patients' SD. Also, it shows that tumor intensity (pixel value) information falls within this range, reducing the SD. It signifies that the intensity of pixels of a tumor in an MRI plays a significant role. Both Aboussaleh *et al* and Bae *et al* mention this fact [51–53].

The second most crucial feature is *Age*. From figure 3, it is clear that samples with the lower feature value of age have positive SHAP values. In other words, the lower age increases the SD of patients. This observation aligns with medical science inference, i.e. the age of GBM patients is crucial in determining SD, i.e. the lower the age, the more the survivability [54].

The third most crucial feature is the *cent_wb_x* shown in figure 3. It is a location-based feature representing the centroid coordinate of a WT along the X-axis of an MRI image (a physical coordinate). The plot shows that this feature negatively impacts prediction with intermediate and higher feature values. That means the higher feature value is responsible for reducing the SD of patients. Here, the X-axis represents the axial view [55], and higher feature values represent the physical coordinates of the central part of the brain. This plot signifies that tumors in the brain's central and latter-mid parts will reduce patients' SD [56]. Similar resemblance can be observed for *cent_at_x* and *cent_nec_x* features, which are centroid of active tumor and centroid of necrosis, respectively.

The fourth most important feature is the *log-sigma-2-0-mm3D_firstorder_Kurtosis (LFK)*. It is a first-order radiomic feature extracted using a LOG filter, which signifies the distribution of voxels without considering their spatial relations [57]. This feature measures the tailedness (outliers) of data distribution. From the plot, we can observe that low kurtosis values are increasing SD, and higher kurtosis values are reducing the SD of patients. Most samples fall within the SHAP- value range of -20 to 60.

The fifth most crucial feature is *log-sigma-2-0-mm-3D_glcm_Correlation*. It is a second-order radiomic feature extracted using a LOG filter, which measures the inter-relationship of intensity between neighbor voxels [57]. The plot shows that higher feature values are responsible for increasing SD, and lower feature values reduce SD. In other words, the higher correlation between voxels value increases SD, and low correlation values reduce SD.

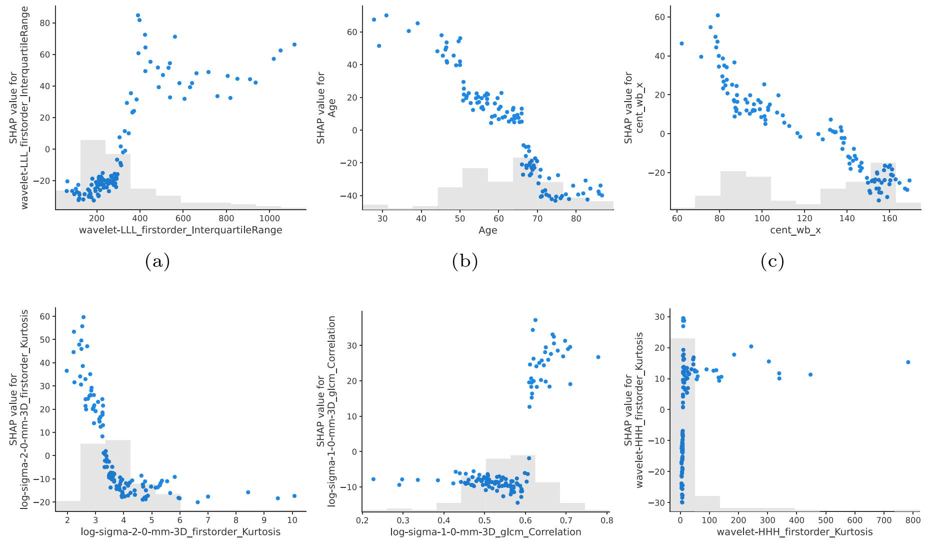
The sixth most important feature is *wavelet-HHH_firstorder_Kurtosis*. It is a first-order radiomic feature extracted using a wavelet filter that uses high-pass filters in the series of z, y, and x directions. The distribution of voxels is independent of their spatial relations, similar to the fourth most important feature. Here, the plot shows that lower feature values are responsible for increasing SD (for more information, see figure 4 for SHAP-distribution plot).

In summary, comparing all the features, we can say that the range of SHAP values for all the features is -40 to $+40$ (in X-axis). Also, with the decreasing of feature importance, the range of SHAP value decreases. That means the features with a low SHAP range have a lower impact on the SD.

Note: most samples and their SHAP-value can also be verified through figure 4, which shows the respective features' SHAP-value and feature value distributions.

Again the SHAP-waterfall plot provides the visual interpretation of features contribution for a single prediction. Figure 5 is a SHAP-waterfall plot for a single sample. The average SD is shown on the X-axis, and the features are arranged on the Y-axis in descending order according to their SHAP values (from top to bottom). From this plot we can analyze, how much the features are impacting negatively (blue) or positively (red) and thereby shift the prediction from the expected outcome $E[f(x)]$. The expected outcome is the average of all the outcomes for all the samples. We observe that for our example sample (for which the plot is generated), the model output is $f(x) = 331.732$. The expected output is $E[f(x)] = 478.91$. This deviation in the model outcome can be understood by quantifying the influences of each of the features.

The SHAP value of each feature in figure 5 depicted this quantification. By adding all of the SHAP values from each feature, it is possible to determine how much each feature (N) contributed to the model output. This is given by $f(x) = E(f(x)) + \sum_N \text{SHAP}$. Here the $\sum_N \text{SHAP}$ represents the sum of the SHAP value of all the features. From this analysis also we can see that the feature *Age* is having a higher impact on the model outcome. For this sample, the Age value is 71.37 and it reduces the average SD by 39.33 days (− (minus) value indicates a reduction in SD). Similarly, *cent_wb_x* value is 164.651, which is also reducing SD by 28.22 days. Whereas mapping *Age, cent_wb_x* features to SHAP-summary Plot (figure 5) or SHAP-distribution plot figure A3 which shows global impacts. We can extract similar observances of reducing SD for these features. For e.g. visualizing *Age, cent_wb_x* feature on SHAP-summary Plot, which shows a higher value of these



(a) (b) (c)

(d) (e) (f)

Figure 4. SHAP value and its distribution for the top six dominant features, say (a) *WIR*, (b) *age*, (c) *cent_wb_x*, (d) *log-sigma-2.0-mm-3D_firstorder_Kurtosis*, (e) *log-sigma-1-0-mm-3D_glcm_Correlation*, (f) *wavelet-HHH_firstorder_Kurtosis*. X-axis displays the feature value of respective features whereas Y-axis displays the SHAP-value of the respective feature. Each blue dot indicates a specific sample. And bars show histograms of feature value. The rest of the features SHAP value distribution can be found in supplementary figure A3.

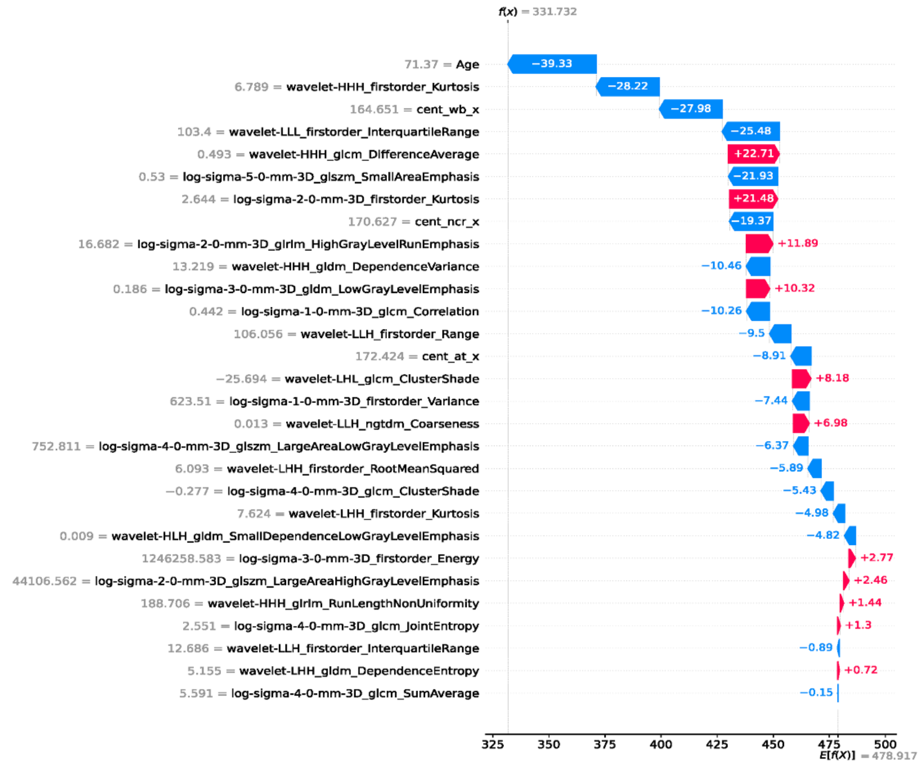


Figure 5. SHAP-waterfall plot of dominant features. This plot shows the features' impact on a particular instance. Each row in the figure shows how the negative (blue) or positive (red) contribution of each feature shifts the value from the expected output ($E[f(x)]$) to the model predicted value ($f(x)$). $f(x) = E(f(x)) +$ shapley value of each feature. Value besides every feature row shows their respective observed value.

features are reducing SD. Similarly, visualizing *Age, cent_wb_x* feature on SHAP-distribution plot figure A3 also shows a reduction in SD. This proves that features show the same behavior both globally and locally.

Further, more information was derived by combining SHAP summary (figure 3), SHAP-distribution plot (figure 4), and PDP plots of the top six dominant features (figure 6) which is explained in section 3.3.2.

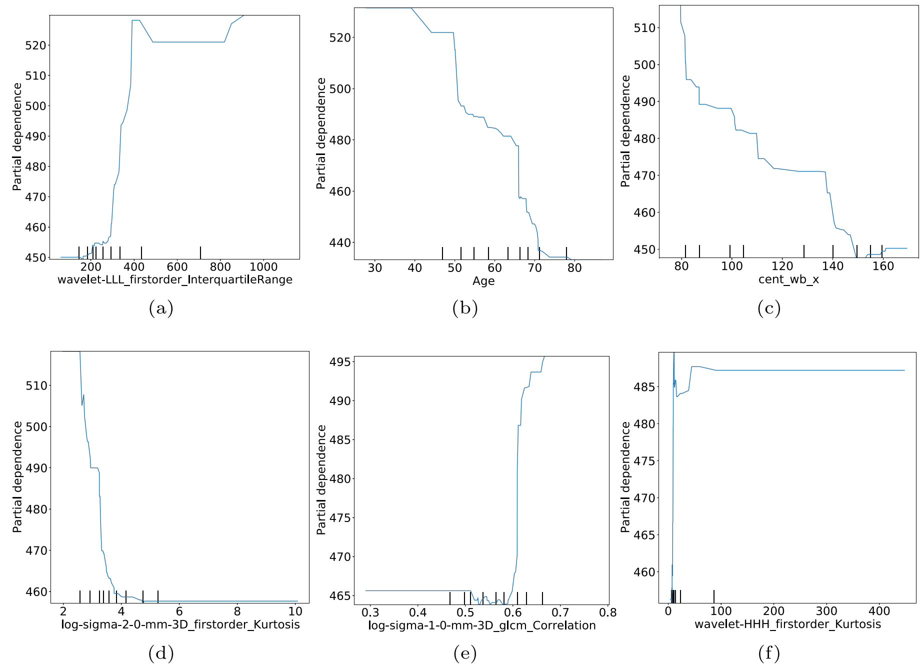


Figure 6. PDP analysis of the top six features. This plot depicted the marginal effect of the (a) *WIR*, (b) *Age*, (c) *cent_wb_x*, (d) *log-sigma-2.0-mm-3D_firstorder_Kurtosis*, (e) *log-sigma-1-0-mm-3D_glcm_Correlation*, (f) *wavelet-HHH_firstorder_Kurtosis* on the survival days. Here, X-axis shows values of the respective feature, whereas Y-axis shows the average rate of change (marginal-impact) respective feature value creates on the target feature. The vertical bar on the X-axis shows most of the samples' distribution. The rest of the features for PDP can be found in supplementary figure A2.

3.3.2. PDP analysis

A PDP shows a marginal effect between desirable and target features (SD) [30]. It shows how a dependent variable changes when an explanatory variable changes, provided all other variables remain constant. If changing the value of a particular feature creates more variation in the average SD indicates that the feature is crucial. In this analysis, we consider the top six features according to their importance (with respect to their absolute SHAP value). These are *WIR, Age, cent_wb_x, LFK, log-sigma-1-0-mm-3D_glcm_Correlation, and wavelet-HHH_firstorder_Kurtosis*. The PDPs of the dominant six features are shown in figure 6 (and plots of the rest of the features are shown in figure A2 supplementary section). The PDPs were arranged in the order of importance (higher to lower) obtained through the SHAP summary plot (figure 3).

Furthermore, visualizing the PDPs, we found the marginal impacts are in line with the order of importance of features that supplements SHAP-value analysis. The detailed analysis of the top six features is: the marginal effect of *WIR* feature on the SD is shown in figure 6. The trend shows value sharply increases within the range 100 to 300, reduces within the 300–350 range, and remains saturated within 350–800 intensity value. It indicates that intensity heterogeneity is very high in the range from 100 to 300, which causes a sharp increase in marginal impact. This suggests that intra-tumor tissues are highly heterogeneous. Comparing PDP (figure 6(a)) with SHAP (figure 3) and its distribution plots (figure A3(a)), we can conclude that this intensity range between 100 and 300 is decreasing SD (testified through a decrease in SHAP value). Hence we can conclude that tumor pixel intensity within this range is detrimental to a patient's survival. Also, as mentioned in this study, higher tumor heterogeneity is associated with increased malignancy [58]. This also complies with the other studies, which suggest that wavelet filters help capture enhanced texture features [58, 59].

Similarly, from the PDP of the *Age* feature (figure 6(b)), we can observe the trend of marginal effect, which shows a maximum deviation in marginal effects for lesser Age patients, signifies maximum impacts on SD. Further, the marginal effect reduces with the increasing Age of patients. Comparing SHAP summary and SHAP-distribution plots, we can observe that after 60 years of Age, there is a decrease in SD (as there is a decrease in SHAP-value beyond this range). Whereas, the PD plot for the *cent_wb_x* feature (shown in figure 6(c)) is the physical coordinate of the WT. The plot shows that the marginal effect is more significant if the centroid is within the range of value, i.e. 75–112 (approx.), and less significant for the 113–160 range of value. Also, comparing these ranges to the SHAP distribution plot, we can observe the former range of values is increasing the SD and the latter is reducing the SD, which signifies tumor lesions in the central or latter part are detrimental to patients.

At the same time, the *log-sigma-2.0-mm-3D_firstorder_Kurtosis* feature is a radiomic first-order statistical information that measures the peakedness of data distribution. For a normal distribution, kurtosis (k) is 3. If $k > 3$, the dataset tends to have significant outliers. If $k < 3$, the dataset has fewer or no outliers. PD plot (figure 6(d)) shows for $k = 3$; there is a higher marginal impact on SD. Comparing PDP with SHAP and SHAP-distribution plots, we can conclude that, for most samples, k is 3, and it increases SD. At the same time, there are enough samples with $k > 3$, decreases SD. It signifies that there are considerable amounts of outliers or intra-heterogeneity among samples. As mentioned by Steven *et al* [60], diffusion kurtosis imaging works on a similar principle of capturing non-normal distribution behavior, which signifies tissue heterogeneity. It is observed that the SD are positively skewed [61].

The *log-sigma-1-0-mm-3D_ glcm_Correlation* is a radiomics feature that calculates the joint likelihood of occurrence of the given pixel pairs with the specified intensity value. At the same time, the gray-level co-occurrence matrix explores spatial relationships between pixels at a specific distance and direction. From the PDP (figure 6(e)), we can observe that if the pixel pairs correlation is more than or equal to 0.6 value, it impacts the SD more. Similarly, observing the correlation threshold of 0.6 in the SHAP and SHAP-distribution plot, one can observe that it impacts positively (having a positive SHAP value). Also, it is mentioned by Sanghani *et al* in their study [62], which shows texture features played a crucial role in SD prediction.

Further, *wavelet-HHH_firstorder_Kurtosis* is a first-order statistical feature like *log-sigma-2.0-mm-3D_firstorderKurtosis* (the fourth most feature), but it is extracted using wavelet filters. Comparing their PDPs (figures 6(d) and (f)) shows both capture kurtosis information but in different dimensions. From the PDPs (shown in figure 6(f)), we can observe that the kurtosis value is steeply increasing between 0 and 100 and then stagnant for the rest of the values. Further, observing SHAP- distribution, we can conclude that most samples are in this range, and samples near $1 - 10$ values are decreasing the SD, and the rest are increasing the SD. However, some samples are sparsely distributed, signifying that they are outliers.

All these signify the importance of these features in determining the SD. With the above analysis, we find the *Age, WIR, cent_wb_x, LFK, log-sigma-1-0-mm-3D_glcm_Correlation, and wavelet-HHH_ firstorder_Kurtosis* plays a crucial role in determining a patient's SD. Similarly, we can analyze other remaining features. Finally, we agree that the *WIR* feature could tell us about tumor heterogeneity associated with high malignancy. Again, the Age feature showed us the trend of survivability, where the survival chances decrease with the patient's increasing age (this is further validated by the KM [34] plot as shown in the supplementary figure A4). At the same time, the centroid of tumors enabled us to locate tumors in the central or latter-central part, which are detrimental for patients. All these analyses using the SHAP and PDPs are analogous to medical findings and related studies. This signifies the model's reliability and validates the explainability methods such as SHAP and PDP.

4. Limitations of the proposed approach and future prospect

The SHAP and PDP techniques are the post-hoc methods that interpret the model after the completion of training. However, for the further understanding of a model, the study of the intrinsic characteristics may help to an extent. Functional imaging like positron emission tomography (PET), functional magnetic resonance imaging (FMRI), and magnetic resonance spectroscopy (MRS) can provide insights into GBM by capturing molecular or physiological information not captured by normal MRI or CT Scans. The methods like the neural ordinary differential equation model (NODE) can provide the learning behavior of a model, especially to understand the spatiotemporal deep feature extraction of a segmentation model [63]. Further, the diffusion imaging modalities such as diffusion kurtosis imaging [64] may help us to understand the underline biological and pathological characteristics of GBM. However, these kind of functional imaging are more complex to analyze, has a high variability across imaging sessions, are more susceptible to noise, and are also expensive. In short, they face several challenges for routine GBM prognosis [65, 66]. Still we believe, integration of these modalities with conventional MRI techniques will enhance the understanding of GBM with added model transparency and interpretability.

5. Conclusion

We have proposed an end-to-end approach for the SD prediction task. We have identified the 29 most dominant features that help predict SD accurately. Again, we validate the optimality of these features using correlation and histogram plots. The trained model performs better on multiple performance metrics. Also, it predicts a more accurate SD than the top-ranking method of the BraTS-2020 competition. Further, we also explore the interpretability of the model to understand its decision globally and locally using post-hoc methods, i.e. SHAP and PDP. Observing these plots, we found that first-order statistical features, Age, location-based and texture features play a crucial role in prediction. Also, these interpretability methods can provide valuable insights into the model that can give human-understandable inferences. The inferences obtained for six dominant features using these interpretability methods were in line with medical facts. We also find that WIR, Age, and location-based features influence the most in predicting SD. We further verify this conclusion using the KM estimation method on the metadata available with the BraTS dataset. Thus, the model is robust in predicting brain tumor patients' survivability. In addition, the interpretability methods can help us to understand model behavior at multiple levels. This will ultimately help to develop trust between medical experts and ML models and incorporate it into clinical practices.

Data availability statement

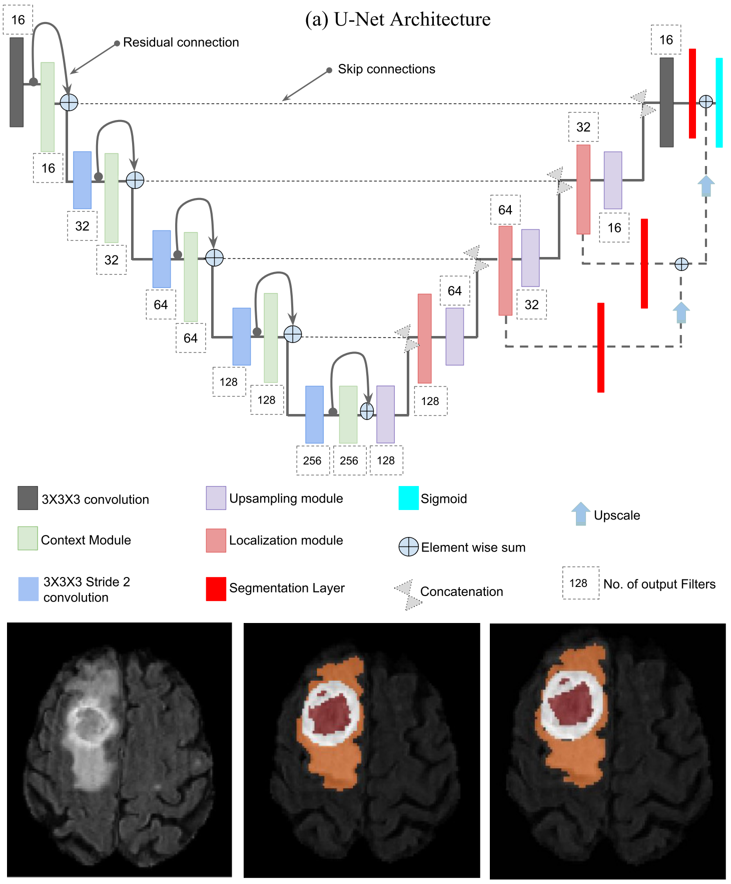
The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Declarations

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. $GUJCOST/STI/2021 - 22/3873ofGUJCOST$, Govt. of Gujarat, India. M S Raval acknowledges Grant No. $GUJCOST/STI/2021 - 22/3858$ of Gujarat Council of Science and Technology (GUJCOST), Govt. of Gujarat, India for computing facility.

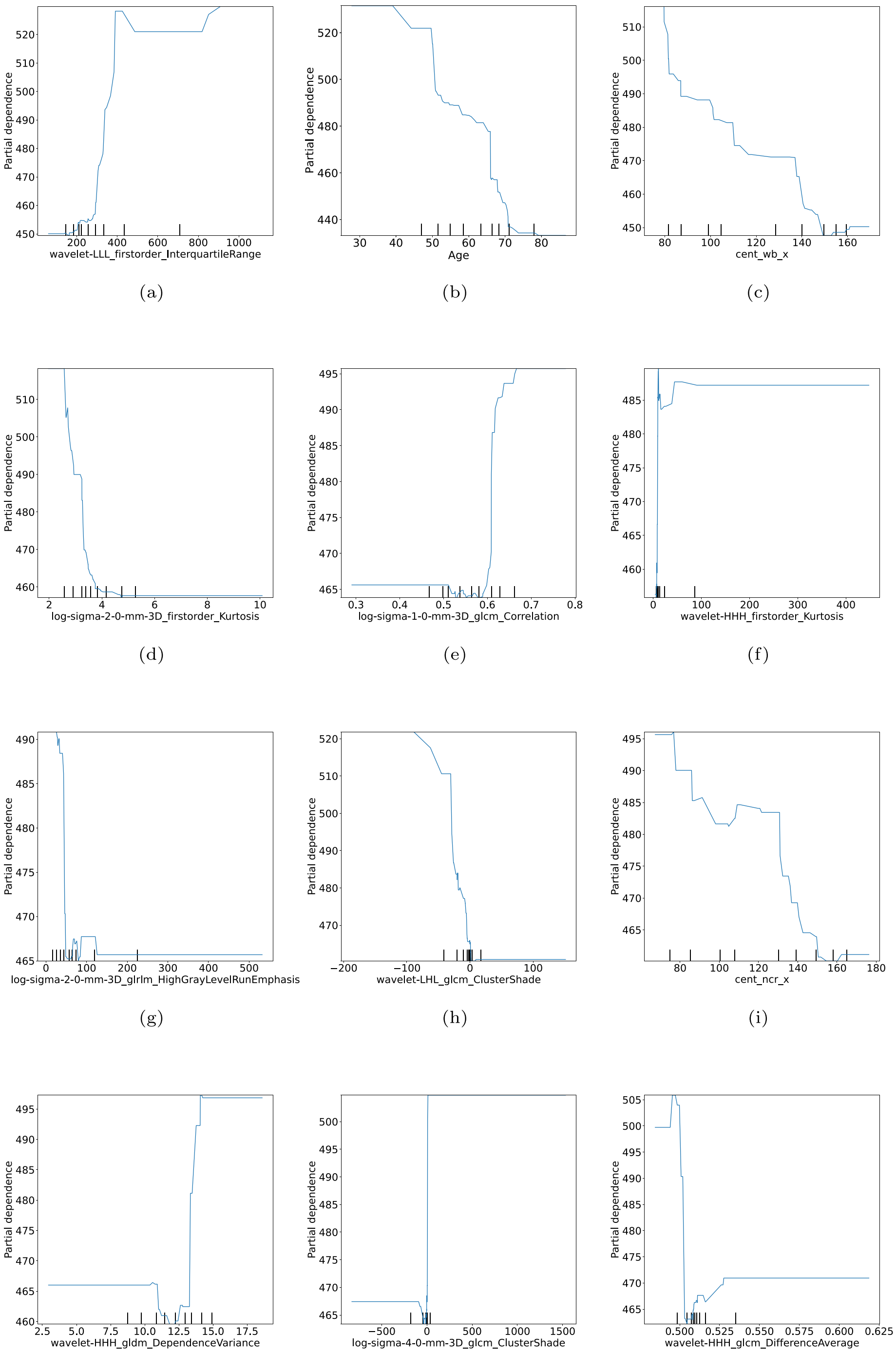
Appendix. Supplementary

Supplementary figures



(b) Input axial-FLAIR (c) Ground truth (d) Segmented result

Figure A1. *No-new network* architecture and segmentation results from that architecture (using BraTS-2020 dataset), (a) 3D U-Net for brain tumor segmentation, (b) input axial-FLAIR slice, (c) axial-ground truth, and (d) axial-segmentation result.



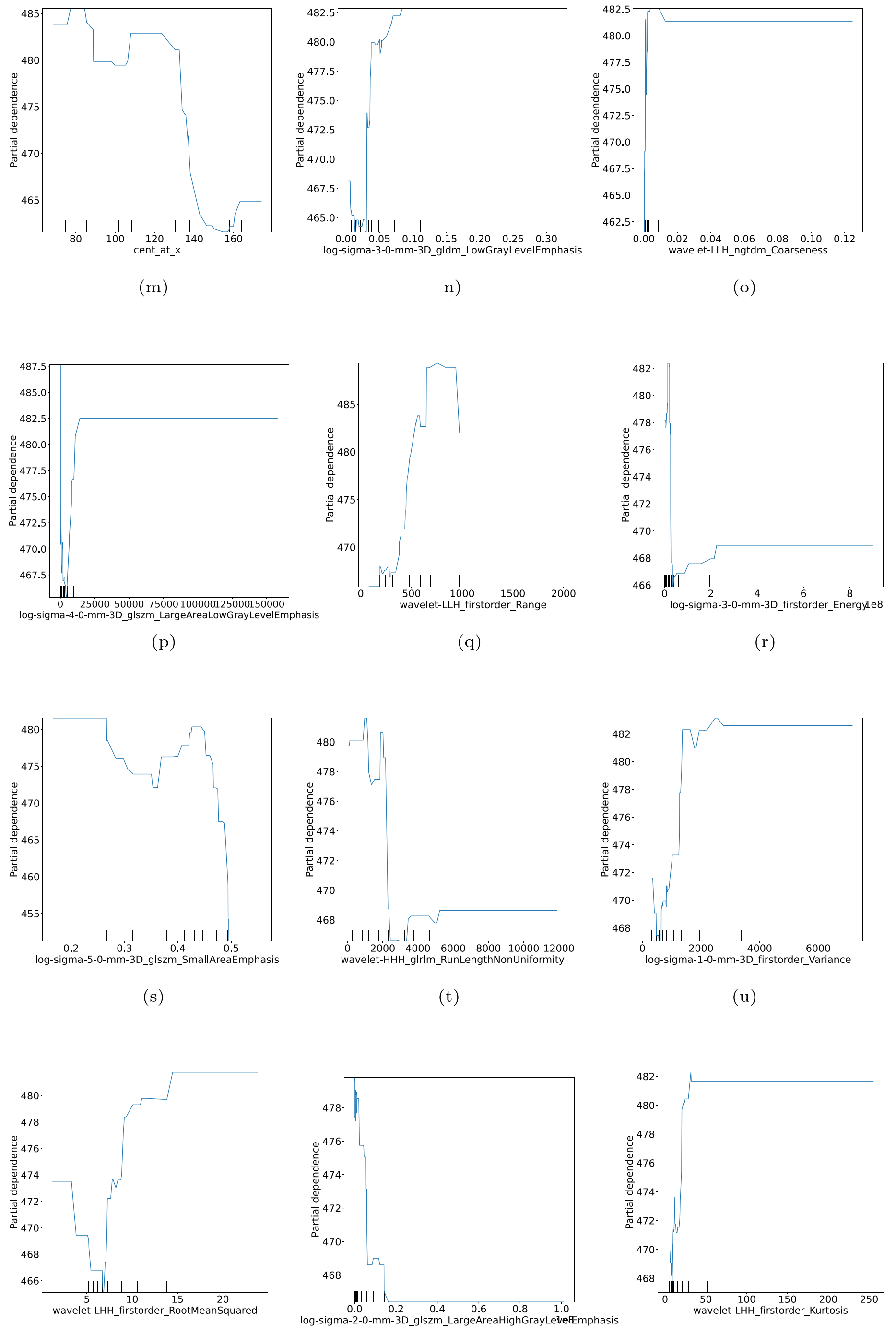
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure A2. PDPs of dominants features: X-axis shows values of respective features, and Y-axis shows the average rate of change of feature effect on target feature. The vertical bars on X-axis show data distribution. This captures global trends of desirable features on the target variable by considering all the samples.



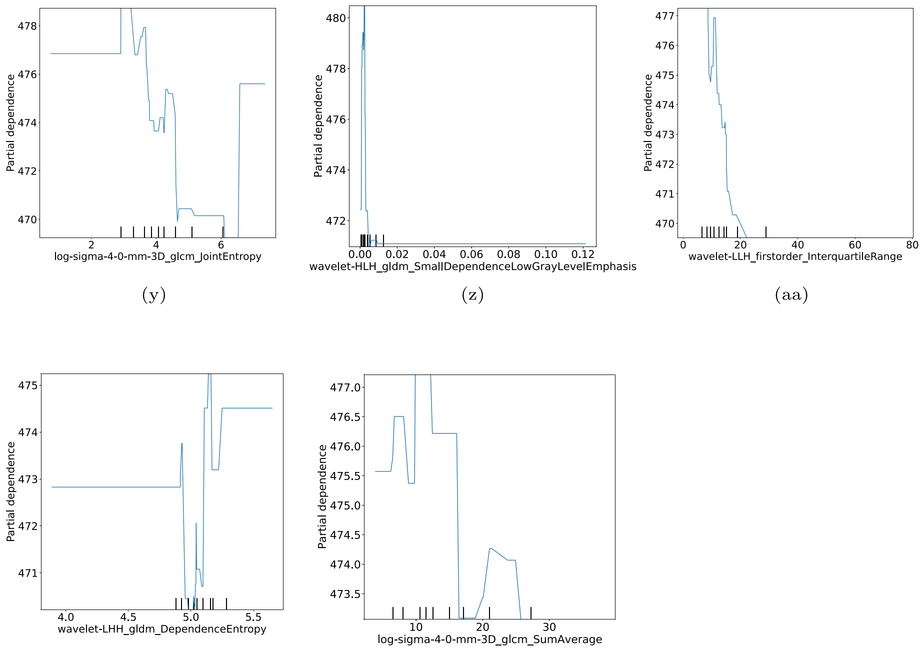
(m) n) (o)

(p) (q) (r)

(s) (t) (u)

(v) (w) (x)

Figure A2. (Continued.)



(y)

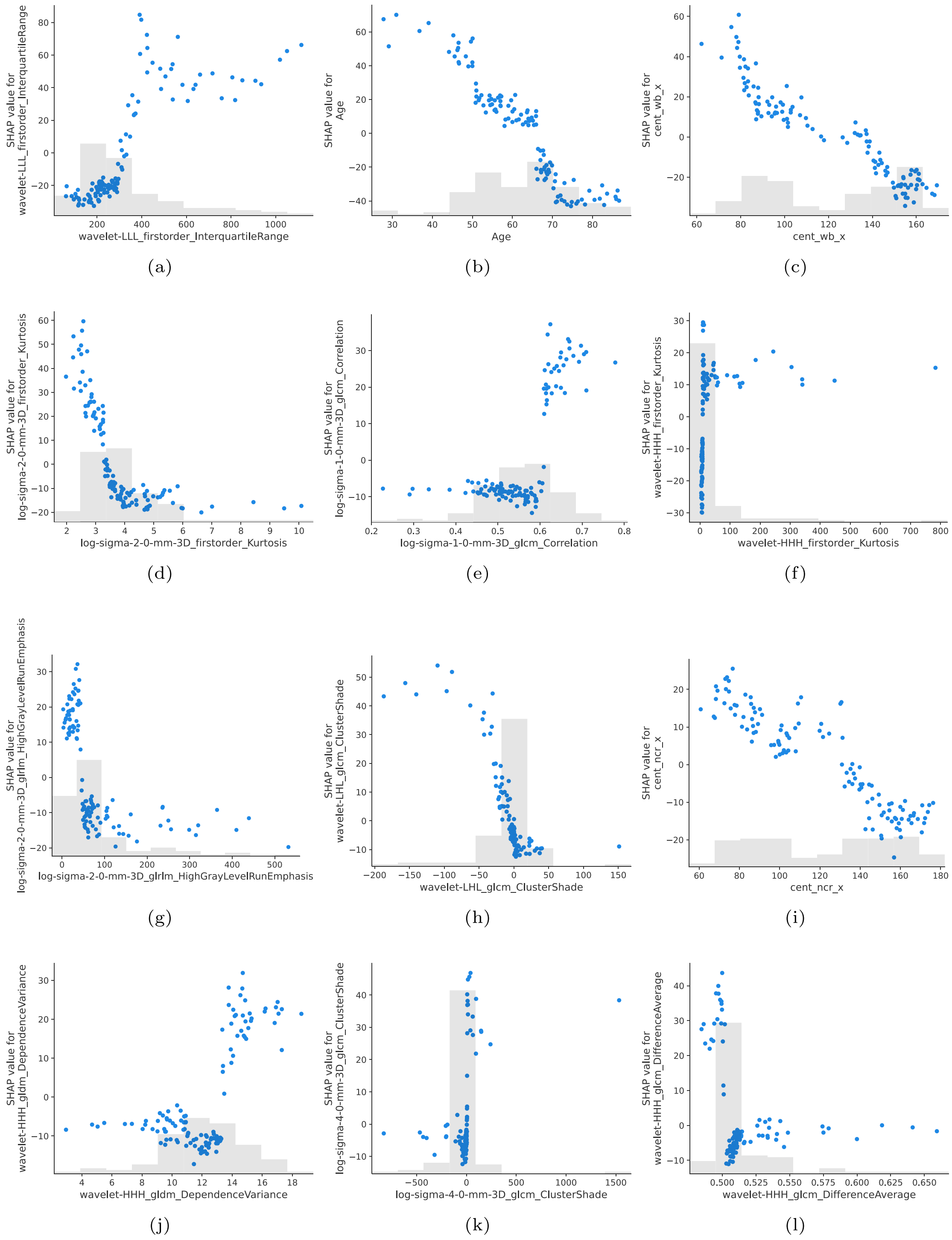
(z)

(aa)

(ab)

(ac)

Figure A2. (Continued.)



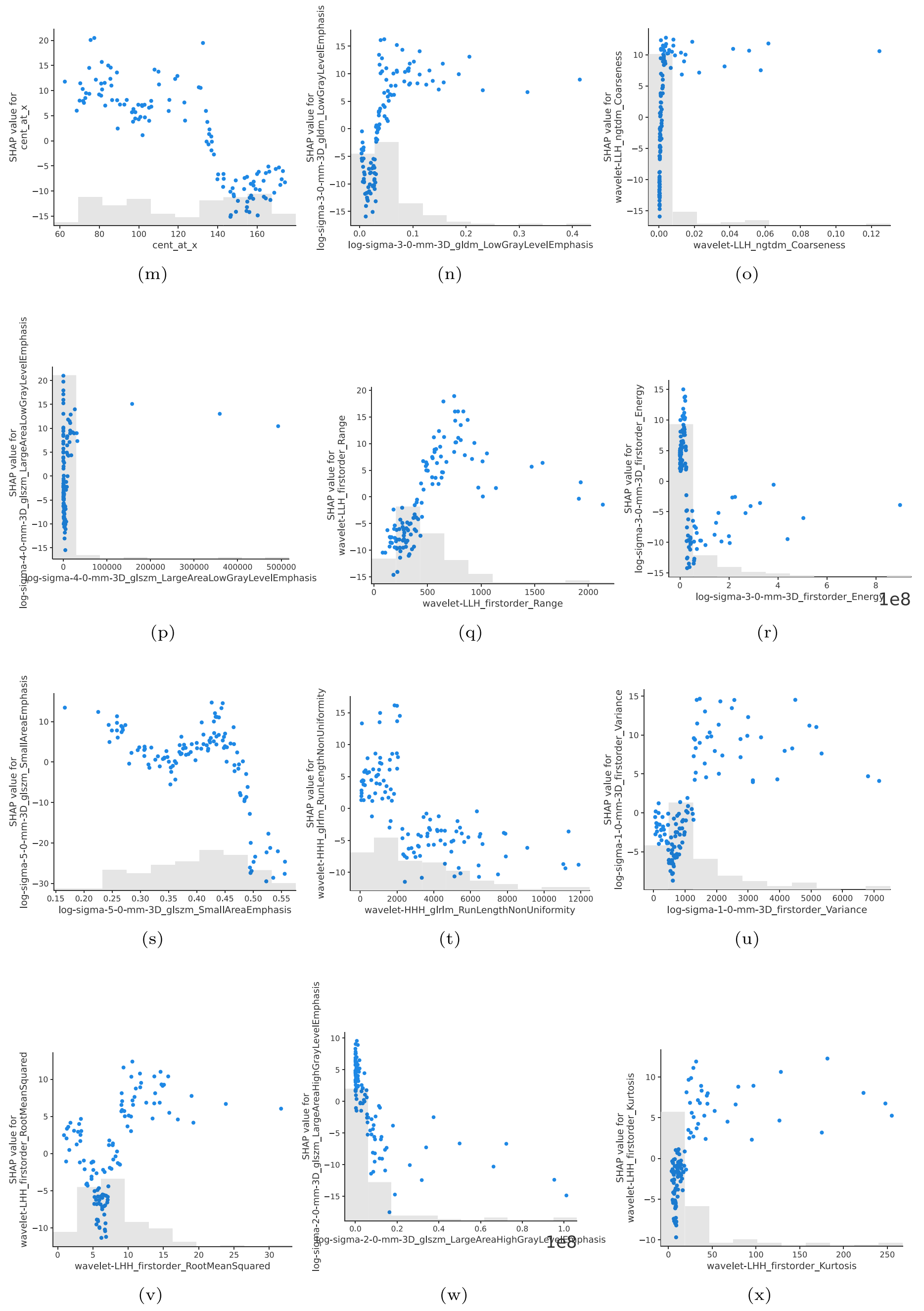
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure A3. SHAP value and its distribution for the dominant features. The X-axis shows the feature value of the respective feature, whereas Y-axis shows the SHAP-value of respective instances. The shaded region shows the distribution of instances. Each dot is an instance from the training dataset. This can help us to visualize and analyze where the majority of SHAP feature value lies, how individually the instances impact target features (range of impact instance wise), and its distribution. This testifies how these values play a role in defining the important feature. From all the SHAP plots, we can observe the magnitude of the SHAP value reduces with the order of importance of features (high to low). *Note:* SHAP value calculates how much feature value changes the model's predicted value from the average.



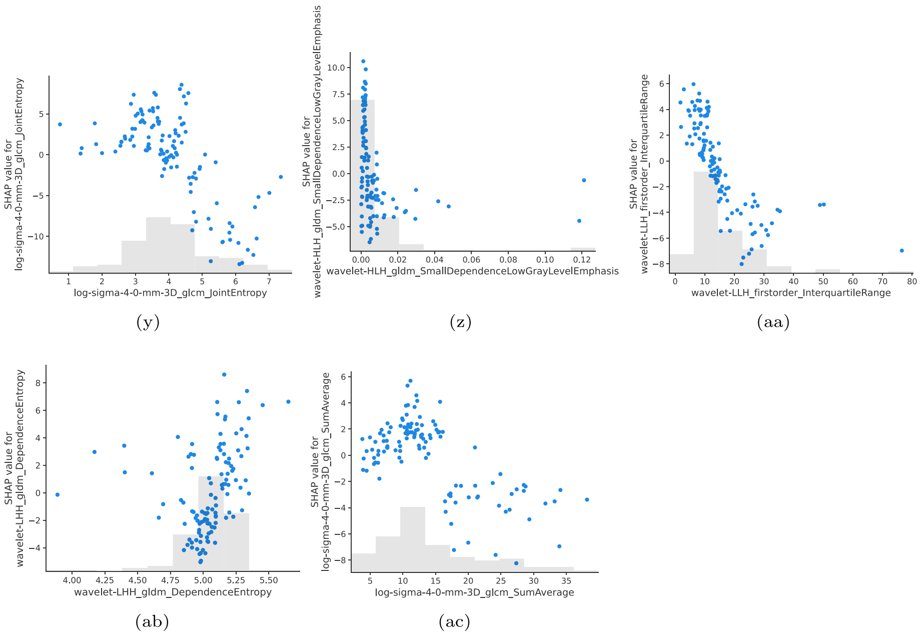
(m) (n) (o)

(p) (q) (r)

(s) (t) (u)

(v) (w) (x)

Figure A3. (Continued.)



(y) (z) (aa)

(ab) (ac)

Figure A3. (Continued.)

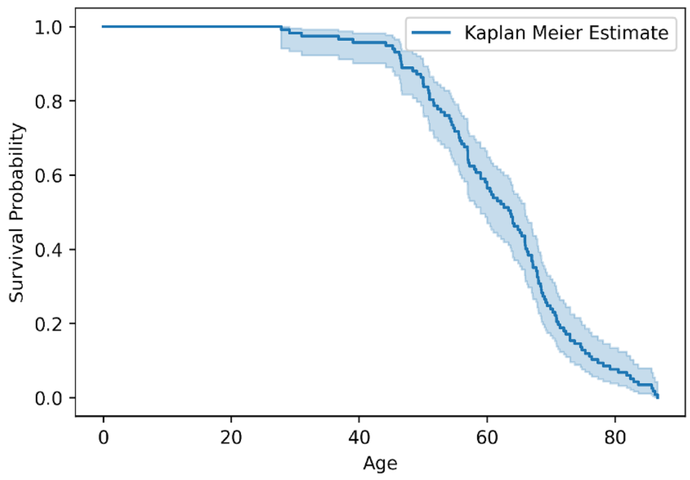


Figure A4. Kaplan–Meier survival plot for survival probability.

The KM estimator measures the percentage of patients that have survived over a certain period after the treatment or surgery. It computes probabilities of the occurrence of events for a duration of time by dividing them into small intervals and re-estimates the probabilities to get the final estimate. The survival probability is computed as follows:

$$S_{t+1} = S_t \times ((N_{t+1} - D_{t+1})/N_{t+1}) \tag{A1}$$

where, N denotes the number of people at risk and D denotes the number of people who died and t is the time interval. The KM survival curve is shown in figure A4. It is a cumulative measure, and the survival remains the same until another individual encounters the risk. From this plot, we observed that the survival probability of older patients is low. The survivability reduces almost linearly after the age of 50 and almost exponentially after the age of 70 and it is very low beyond the age of 80. This KM analysis on the metadata supports the PDP analysis, which shows the exponential decay of survivability from the age near 70.

Supplementary tables

Table A1. Dominant features obtained through RFE and their details.

Features with their descriptions
Age: age information given with the dataset.
cent-at-y : centroid of active tumor across the y axis.
pos-ed-wb-x: centroid of enhancing tumor w.r.t brain centroid across x axis.
original-shape-LeastAxisLength: it calculates smallest axis length of the ROI.
Wavelet-LLH-firstorder-Maximum: it measures maximum gray intensity within the ROI after applying the Wavelet LLH-band filter.
Wavelet-LLH-Gldm-DependenceVariance: it calculates variance in the dependence matrix of the image after applying the Wavelet LLH-band filter.
Wavelet-LLH-Glrlm-LongRunLowGrayLevel-Emphasis: it estimates the length in the terms of successive pixels run lengths with lower gray-level intensity values, after applying Wavelet transform.
Wavelet-LHL-Glcm-Correlation: it assesses the correlation between gray-level values and their related voxels in the gray-level co-occurrence matrix, after applying the Wavelet LHL band filter.
Wavelet-LHH-Gldm-DependenceNonUniformityNormalized: it assesses the similarity between dependencies in an image, after applying the Wavelet LHH band filter.
Wavelet-LHH-Gldm-SmallDependenceHighGrayLevelEmphasis: it assesses the combined distribution of small dependence with higher gray-level values after applying the Wavelet LHH band filter.
Wavelet-LHH-Glszm-ZoneEntropy: it evaluates the randomization of zone sizes and gray level values in the distribution after applying the Wavelet LHH band filter.
Wavelet-HLL-Glcm-Imc1: it assesses the correlation between the probability distribution of two pixels after applying the Wavelet HLL band filter.
Wavelet-HLH-firstorder-Kurtosis: it assesses the peakedness of the spread of pixel intensities in the given image after applying the Wavelet HLH band filter.
Wavelet-HLH-Gldm-DependenceEntropy: it assesses randomness in the dependencies of an image after applying the Wavelet HLH band filter.
Wavelet-HLH-Gldm-SmallDependenceLowGrayLevel-Emphasis: it assesses the combined distribution of small dependence with higher gray-level values after applying Wavelet HLH band filter.
Wavelet-HHH-Glcm-MaximumProbability: it finds the most frequently occurring neighboring pair of intensity values from the grey-level co-occurrence matrix after applying the Wavelet HHH band filter.
Wavelet-LLL-Glcm-Correlation: it assesses the association between pairs and their corresponding voxel intensity value after applying the Wavelet LLL band filter.
LoG-sigma-1-0-mm-3D-Glcm-Correlation: it assesses the association between pairs and their respective voxel intensity value after applying LoG filter with sigma value 1.
Wavelet-LLH-Ngtdm-Strength: it assesses strength in an image after applying a Wavelet filter using the LLH band.
LoG-sigma-5-0-mm-3D-Glrlm-RunLengthNonUniformity-Normalized: It assesses the homogeneity in the gray level run lengths in the image after applying the LoG filter with sigma value 5.
LoG-sigma-3-0-mm-3D-Glrlm-RunVariance: it calculates variance in the gray-level run-lengths in the image, after applying LoG filter with sigma value 3.
LoG-sigma-2-0-mm-3D-Glcm-ClusterShade: it calculates uniformity in the gray-level co-occurrence matrix after applying the LoG filter with sigma value 2.
LoG-sigma-5-0-mm3D-firstorder-TotalEnergy: it assesses the localized change of the image after applying the LoG filter with sigma value 5.
LoG-sigma-3-0-mm-3D-Glcm-MaximumProbability: it assesses the occurrences of the most prevalent pairing of neighboring intensity values in the grey-level co-occurrence matrix after applying the LoG filter with sigma value 5.
LoG-sigma-2-0-mm-3D-firstorder-90Percentile: it assesses 90th percentile intensity values of an image after applying LoG filter with sigma value 2.
LoG-sigma-2-0-mm-3D-firstorder-Skewness: it calculates the asymmetry of the distribution of intensity values that deviates from the mean intensity value after applying the LoG filter with sigma value 2.
LoG-sigma-1-0-mm-3D-Glcm-MCC: it assesses the complexity of the texture in the co-occurrence matrix of an image after applying the LoG filter with sigma value 1.
Wavelet-LLL-Glszm-SmallAreaEmphasis: it assesses the number of connected voxels with the same gray-level intensity value or the spread of smaller size zones after applying Wavelet LLL band filter.
Wavelet-HLL-Glcm-MCC: it assesses the complexity of the texture in the co-occurrence matrix of an image after applying the Wavelet HLL band filter.

Table A2. Dominant feature set through PI and their weights. The threshold value of the weights is 100.

Weight	Features
1309.94	***Age*** : age information given with the dataset.
0761.33	***LoG-sigma-1-0-mm-3D-glcm-Correlation***: it assesses the association between pairs and its respective voxel intensity value after applying the LoG filter with sigma value 1.
0722.95	***Wavelet-HHH-Gldm-DependenceVariance***: it calculates variance in the dependence matrix of the image after applying the Wavelet HHH band filter.
0678.10	***LoG-sigma-4-0-mm-3D-Glcm-JointEntropy***: it calculates the randomness in neighborhood intensity values.
0669.58	***LoG-sigma-2-0-mm-3D-firstorder-Kurtosis***: it assesses the peakiness of the intensity distribution of a given image after applying the LoG filter with sigma value 2.
0558.74	***LoG-sigma-2-0-mm-3D-Glrlm-HighGrayLevelRunEmphasis***: it assesses the spread of the image's upper gray-level values in the image.
0555.77	***Wavelet-HLH-Gldm-SmallDependence-LowGrayLevelEmphasis***: it assesses the combined spread of small-dependence with lower gray-level values after applying Wavelet HLH band filter.
0509.37	***LoG-sigma-3-0-mm-3D-Gldm-LowGrayLevelEmphasis***: it calculates the spread of low gray-level values in the image.
0476.60	***cent-ncr-x***: centroid of necrosis across *x*-axis.
0464.70	***Wavelet-LLL-firstorder-InterquartileRange***: it assesses the difference between the 75th and 25th percentile of the image array after applying the Wavelet LLL band filter.
0444.84	***LoG-sigma-4-0-mm-3D-Glcm-ClusterShade***: it calculates uniformity in the gray level co-occurrence matrix after applying LoG filter with sigma value 4.
0438.99	***Wavelet-LHH-firstorder-RootMeanSquared***: it assesses the root-mean-square of the intensity value of an image after applying the Wavelet LHH band filter.
0420.35	***LoG-sigma-4-0-mm-3D-Glcm-SumAverage***: it assesses the relationship between pair occurrences with lower intensity values and pair occurrences with higher intensity values, after applying LoG filter with sigma value 4.
0406.08	***Wavelet-HHH-Glrlm-RunLengthNonUniformity***: it assesses the homogeneity between different run lengths of the image.
0395.63	***LoG-sigma-5-0-mm-3D-Glszm-SmallAreaEmphasis***: it assesses the spread of small size-zones or the number of connected voxels that have the same gray-level intensity value, after applying LoG filter with sigma value 5.
0357.96	***Wavelet-LLH-Ngtdm-Coarseness***: It assesses the spatial rate of change in the intensity value after applying the Wavelet LLH band filter.
0357.51	***Wavelet-LLH-firstorder-InterquartileRange***: it assesses the difference between the 75th and 25th percentile of the image array after applying the Wavelet LLH band filter.
0340.11	***cent-at-x***: centroid of active tumor across *x*-axis.
0314.18	***LoG-sigma-4-0-mm-3D-Glszm-LargeAreaLowGrayLevel-Emphasis***.
0282.22	***Wavelet-HHH-firstorder-Kurtosis***: it assesses the peakedness of the spread of the image's intensity values after applying the Wavelet HHH band filter.
0247.80	***Wavelet-HHH-Glcm-DifferenceAverage***: it assesses the relationship between the occurrences of pairings with similar intensity values and those with different intensity values after applying the Wavelet filter.
0247.36	***cent-wb-x***: centroid of whole-tumor brain across *x*-axis.
0232.56	***LoG-sigma-3-0-mm-3D-firstorder-Energy***: it assesses the magnitude of voxel values in an image.
0229.91	***LoG-sigma-1-0-mm-3D-firstorder-Variance***: it measures the distribution spread about the mean intensity value after applying LoG filter with sigma value 1.
0226.71	***Wavelet-LHH-firstorder-Kurtosis***: it assesses the image's peakedness in terms of intensity distribution, applying Wavelet LHH band filter.
0217.80	***LoG-sigma-2-0-mm-3D-Glszm-LargeAreaHighGrayLevel-Emphasis***: it assesses the combined spread of larger size-zones with higher gray-level values, after applying the LoG filter with sigma value 1.
0183.47	***Wavelet-LLH-firstorder-Range***: it assesses the distribution of gray-level values of an image.
0131.10	***Wavelet-LHH-Gldm-DependenceEntropy***: it assesses randomness in the dependencies of an image after applying Wavelet LHH band filter.
0118.90	***Wavelet-LHL-Glcm-ClusterShade***: it calculates uniformity in the gray level co-occurrence matrix after applying the Wavelet LHL band filter.

Table A3. Feature annotation of a correlation matrix.

Index no.	Features name	Features type
1	*Age*	*Meta-Data*
2	*cent-at-x*	*Image-based*
3	*cent-ncr-x*	*Image-based*
4	*cent-wb-x*	*Image-based*
5	*LoG-sigma-1-0-mm-3D-FirstorderVariance*	*Radiomics-based*
6	*LoG-sigma-1-0-mm-3D-Glcm-Correlation*	*Radiomics-based*
7	*LoG-sigma-2-0-mm-3D-FirstorderKurtosis*	*Radiomics-based*
8	*LoG-sigma-2-0-mm-3D-Glrlm-HighGrayLevelRunEmphasis*	*Radiomics-based*
9	*LoG-sigma-2-0-mm-3D-Glszm-LargeAreaHighGrayLevelEmphasis*	*Radiomics-based*
10	*LoG-sigma-3-0-mm-3D-Firstorder-Energy*	*Radiomics-based*
11	*LoG-sigma-3-0-mm-3D-Gldm-LowGrayLevelEmphasis*	*Radiomics-based*
12	*LoG-sigma-4-0-mm-3D-Glcm-ClusterShade*	*Radiomics-based*
13	*LoG-sigma-4-0-mm-3D-Glcm-JointEntropy*	*Radiomics-based*
14	*LoG-sigma-4-0-mm-3D-Glcm-SumAverage*	*Radiomics-based*
15	*LoG-sigma-4-0-mm-3D-Glszm-LargeAreaLowGrayLevelEmphasis*	*Radiomics-based*
16	*LoG-sigma-5-0-mm-3D-Glszm-SmallAreaEmphasis*	*Radiomics-based*
17	*Wavelet-HHH-Firstorder-Kurtosis*	*Radiomics-based*
18	*Wavelet-HHH-Glcm-DifferenceAverage*	*Radiomics-based*
19	*Wavelet-HHH-Gldm-DependenceVariance*	*Radiomics-based*
20	*Wavelet-HHH-Glrlm-RunLengthNonUniformity*	*Radiomics-based*
21	*Wavelet-HLH-Gldm-SmallDependence-LowGrayLevelEmphasis*	*Radiomics-based*
22	*Wavelet-LHH-Firstorder-Kurtosis*	*Radiomics-based*
23	*Wavelet-LHH-Firstorder-RootMeanSquared*	*Radiomics-based*
24	*Wavelet-LHH-Gldm-DependenceEntropy*	*Radiomics-based*
25	*Wavelet-LHL-Glcm-ClusterShade*	*Radiomics-based*
26	*Wavelet-LLH-Firstorder-InterquartileRange*	*Radiomics-based*
27	*Wavelet-LLH-Firstorder-Range*	*Radiomics-based*
28	*Wavelet-LLH-Ngtdm-Coarseness*	*Radiomics-based*
29	*Wavelet-LLL-Firstorder-InterquartileRange*	*Radiomics-based*

Table A4. Example of calculating SHAP value. Let us consider feature set (F) = {A, B, D}, and values (contribution) of features are: v{A} = 8, v{B} = 10, v{D} = 9, v{A, B} = 18, v{A, D} = 20, v{B, D} = 22 and v{A, B, D} = 25. The bold text shows the arrangement of feature A in the possible combinations of features.

Possible combinations of feature	Marginal combination		
	Feature **A**	Feature B	Feature D
{**A**, B, D}	v{A} -ϕ = 8	v{A, B}-v{A} = 10	v{A,B,D} -v{A,B} = 7
{**A**, D, B}	v{A} -ϕ = 8	v{A, B, D} -v{A, D} = 5	v{A,D} -v{A} = 12
{D, B, **A**}	v{A, B, D} -v{D, B} = 25-22 = 3	v{D, B} -v{B} = 12	v{D} -ϕ = 9
{B, **A**, D}	v{A,B} -v{B} = 8	v{B} -ϕ = 10	v{A,B,D} -v{A,B} = 7
{D, **A**, B}	v{A, D} -v{D} = 11	v{A, B, D} -v{A, D} = 5	v{D}—}ϕ = 9
{B, D, **A**}	v{A, B, D} -v{B, D} = 3	v{B}—ϕ = 10	v{B, D} -v{B} = 12
SHAP value	(8 + 8 + 8 + 10 + 11 + 3) | 6 = 6.833	(10 + 5 + 12 + 10 + 5 + 10) | 6 = 8.667	(7 + 12 + 9 + 7 + 9 + 12) | 6 = 9.334

ORCID iDs

Snehal Rajput https://orcid.org/0000-0001-8240-3740
Rupal A Kapdi https://orcid.org/0000-0003-1995-4149
Mehul S Raval https://orcid.org/0000-0002-3895-1448
Mohendra Roy https://orcid.org/0000-0001-5815-3294

References

[1] Hanif F, Muzaffar K, Perveen K, Malhi S M and Simjee S U 2017 Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment *Asian Pac. J. Cancer Prev.* **18** 3

[2] Ostrom Q T, Cioffi G, Waite K, Kruchko C and Barnholtz-Sloan J S 2021 CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2014–2018 *Neuro-Oncology* **23** iii1–iii105

[3] Rindi G *et al* 2018 A common classification framework for neuroendocrine neoplasms: an international agency for research on cancer (IARC) and World Health Organization (WHO) expert consensus proposal *Mod. Pathol.* **31** 1770–86

[4] Fernández-Llaneza D, Gondová A, Vince H, Patra A, Zurek M, Konings P, Kagelid P and Hultin L 2022 Towards fully automated segmentation of rat cardiac MRI by leveraging deep learning frameworks *Sci. Rep.* **12** 9193

[5] Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Springer) pp 234–41

[6] McKinley R, Meier R and Wiest R 2018 Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation *Int. MICCAI Brainlesion Workshop* (Springer) pp 456–65

[7] McKinley R, Rebsamen M, Daetwyler K, Meier R, Radojewski P and Wiest R 2020 Uncertainty-driven refinement of tumor-core segmentation using 3D-to-2D networks with label uncertainty (arXiv:2012.06436)

[8] Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J S, Freymann J B, Farahani K and Davatzikos C 2017 Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features *Sci. Data* **4** 1–13

[9] Bakas S *et al* 2018 Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge (arXiv:1811.02629)

[10] Menze B H *et al* 2014 The multimodal brain tumor image segmentation benchmark (brats) *IEEE Trans. Med. Imaging* **34** 1993–2024

[11] Rajput S and Raval M S 2020 A review on end-to-end methods for brain tumor segmentation and overall survival prediction (arXiv:2006.01632)

[12] Jia H, Cai W, Huang H and Xia Y 2020 H2nf-net for brain tumor segmentation using multimodal MR imaging: 2nd place solution to brats challenge 2020 segmentation task (arXiv:2012.15318)

[13] Wang Y, Zhang Y, Hou F, Liu Y, Tian J, Zhong C, Zhang Y and He Z 2020 Modality-pairing learning for brain tumor segmentation (arXiv:2010.09277)

[14] McKinley R, Rebsamen M, Meier R and Wiest R 2019 Triplanar ensemble of 3d-to-2d CNNS with label-uncertainty for brain tumor segmentation *Int. MICCAI Brainlesion Workshop* (Springer) pp 379–87

[15] Asenjo J M and Solís A M L 2021 MRI brain tumor segmentation using a 2D-3D U-Net ensemble *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Springer International Publishing) pp 354–66

[16] Myronenko A 2018 3D MRI brain tumor segmentation using autoencoder regularization *Int. MICCAI Brainlesion Workshop* (Springer) pp 311–20

[17] Crimi A and Bakas S 2020 Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries *5th Int. Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019 (Shenzhen, China, 17 October 2019) (Revised Selected Papers, Part I)* vol 11992 (Springer Nature)

[18] Isensee F, Jaeger P F, Full P M, Vollmuth P and Maier-Hein K H 2020 nnU-Net for brain tumor segmentation (arXiv:2011.00848)

[19] Braunstein V 2021 Nvidia data scientists take top spots in miccai 2021 brain tumor segmentation challenge (available at: https://developer.nvidia.com/blog/nvidia-data-scientists-take-top-spots-in-miccai-2021-brain-tumor-segmentation-challenge/)

[20] Agravat R R and Raval M S 2021 A survey and analysis on automated glioma brain tumor segmentation and overall patient survival prediction *Arch. Comput. Methods Eng.* **28** 1–36

[21] Karami G, Giuseppe Orlando M, Delli Pizzi A, Caulo M and Del Gratta C 2021 Predicting overall survival time in glioblastoma patients using gradient boosting machines algorithm and recursive feature elimination technique *Cancers* **13** 4976

[22] Baid U, Rane S U, Talbar S, Gupta S, Thakur M H, Moiyadi A and Mahajan A 2020 Overall survival prediction in glioblastoma with radiomic features using machine learning *Front. Comput. Neurosci.* **14** 61

[23] Hermida L C, Gertz E M and Ruppin E 2022 Predicting cancer prognosis and drug response from the tumor microbiome *Nat. Commun.* **13** 2022

[24] Walid M S 2008 Prognostic factors for long-term survival after glioblastoma *Perm. J.* **12** 45

[25] Feng X, Dou Q, Tustison N and Meyer C 2019 Brain tumor segmentation with uncertainty estimation and overall survival prediction *Int. MICCAI Brainlesion Workshop* (Springer) pp 304–14

[26] Vale-Silva L A and Rohr K 2021 Long-term cancer survival prediction using multimodal deep learning *Sci. Rep.* **11** 2021

[27] Bommineni V L 2021 PieceNet: a redundant UNet ensemble *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Springer) pp 331–41

[28] Isensee F, Kickingereder P, Wick W, Bendszus M and Maier-Hein K H 2017 Brain tumor segmentation and radiomics survival prediction: contribution to the brats 2017 challenge *Int. MICCAI Brainlesion Workshop* (Springer) pp 287–97

[29] Spyridon (Spyros) BCS 2021 Validation survival leaderboard 2020 (available at: www.cbica.upenn.edu/BraTS20//lboardValidationSurvival.html) (Accessed 12 June 2021)

[30] Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Stat.* **29** 1189–232

[31] Friedman J H and Popescu B E 2008 Predictive learning via rule ensembles *Ann. Appl. Stat.* **2** 916–54

[32] Lundberg S M and Lee S I 2017 A unified approach to interpreting model predictions *Proc. 31st Int. Conf. on Neural Information Processing Systems* pp 4768–77

[33] Lundberg S M, Erion G, Chen H, DeGrave A, Prutkin J M, Nair B, Katz R, Himmelfarb J, Bansal N and Lee S I 2020 From local explanations to global understanding with explainable AI for trees *Nat. Mach. Intell.* **2** 56–67

[34] Goel M K, Khanna P and Kishore J 2010 Understanding survival analysis: Kaplan-meier estimate *Int. J. Ayurveda Res.* **1** 274

[35] Van Griethuysen J J, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan R G, Fillion-Robin J C, Pieper S and Aerts H J 2017 Computational radiomics system to decode the radiographic phenotype *Cancer Res.* **77** e104–7

[36] Singh S P and Urooj S 2015 Wavelets: biomedical applications *Int. J. Biomed. Eng. Technol.* **19** 1–25

[37] Kong H, Akakin H C and Sarma S E 2013 A generalized Laplacian of gaussian filter for blob detection and its applications *IEEE Trans. Cybern.* **43** 1719–33

[38] Pedregosa F *et al* 2012 Scikit-learn: machine learning in python (arXiv:2012.06436)

[39] MIT M K and Lopuhin K 1965 permutation_importance (available at: https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html)

[40] Fernáandez-Delgado M, Cernadas E, Barro S and Amorim D 2014 Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15** 3133–81

[41] Puybareau E, Tochon G, Chazalon J and Fabrizio J 2019 Segmentation of gliomas and prediction of patient overall survival: a simple and fast procedure *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ed A Crimi, S Bakas, H Kuijf, F Keyvan, M Reyes and T van Walsum (Cham: Springer) pp 199–209

[42] Agravat R R and Raval M S 2019 Brain tumor segmentation and survival prediction *Int. MICCAI Brainlesion Workshop* (Springer) pp 338–48

[43] Ishwaran H, Kogalur U B, Blackstone E H and Lauer M S 2008 Random survival forests *Annals of Applied Statistics* **2** 841–60

[44] Rajput S, Agravat R, Roy M and Raval M S 2021 Glioblastoma multiforme patient survival prediction (arXiv:2101.10589)

[45] Rozemberczki B, Watson L, Bayer P, Yang H T, Kiss O, Nilsson S and Sarkar R 2022 The shapley value in machine learning (arXiv:2202.05594)

[46] Molnar C 2021 *Interpretable Machine Learning A Guide for Making Black Box Models Explainable* (Leanpub)

[47] Pan X, Zhang T, Yang Q, Yang D, Rwigema J C and Qi X S 2020 Survival prediction for oral tongue cancer patients via probabilistic genetic algorithm optimized neural network models *The British Journal of Radiology* **93** 20190825

[48] Molina G, Chawla A, Clancy T E and Wang J American Society of Clinical Oncology (ASCO) 2019 The correlation between the proportion of patients with pancreatic ductal adenocarcinoma who received neoadjuvant therapy and overall survival between 2004 and 2015 *J. Clin. Oncol.* **37** 395–395

[49] Minoru 2021 Regression—what does the median absolute error metric say about the models? (Version: 13 April 2017) (available at: https://stats.stackexchange.com/q/253892) (Accessed 12 June 2021)

[50] Ali M J, Akram M T, Saleem H, Raza B and Shahid A R 2021 Glioma segmentation using ensemble of 2D/3D U-Nets and survival prediction using multiple features fusion *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Springer) pp 189–99

[51] Aboussaleh I, Riffi J, Mahraz A M and Tairi H 2021 Brain tumor segmentation based on deep learning's feature representation *J. Imaging* **7** 269

[52] Bae S, Choi Y S, Ahn S S, Chang J H, Kang S G, Kim E H, Kim S H and Lee S K 2018 Radiomic MRI phenotyping of glioblastoma: improving survival prediction *Radiology* **289** 797–806

[53] Tessamma T and Ananda Resmi S 2010 Texture description of low grade and high grade glioma using statistical features in brain MRIs (ACEEE) *Int. J. Eng. Technol.* **4**

[54] ASCO ASoCO Brain tumor: statistics (available at: www.cancer.net/cancer-types/brain-tumor/statistics)

[55] Mahmoudzadeh A P and Kashou N H 2014 Interpolation-based super-resolution reconstruction: effects of slice thickness *J. Med. Imaging* **1** 034007

[56] Fyllingen E H, Bø L E, Reinertsen I, Jakola A S, Sagberg L M, Berntsen E M, Salvesen Ø and Solheim O 2021 Survival of glioblastoma in relation to tumor location: a statistical tumor atlas of a population-based cohort *Acta Neurochir.* **163** 1895–905

[57] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti A G and Bellomi M 2018 Radiomics: the facts and the challenges of image analysis *Eur. Radiol. Exp.* **2** 1–8

[58] Gupta M, Rajagopalan V and Prabhakar Rao B V V S N 2019 Glioma grade classification using wavelet transform-local binary pattern based statistical texture features and geometric measures extracted from MRI *J. Exp. Theor. Artif. Intell.* **31** 57–76

[59] Deepa B, Sumithra M, Kumar R M and Suriya M 2021 Weiner filter based hough transform and wavelet feature extraction with neural network for classifying brain tumor *2021 6th Int. Conf. on Inventive Computation Technologies (ICICT)* (IEEE) pp 637–41

[60] Steven A J, Zhuo J and Melhem E R 2014 Diffusion kurtosis imaging: an emerging technique for evaluating the microstructural environment of the brain *Am. J. Roentgenol.* **202** W26–W33

[61] Der G and Everitt B S 2005 Survival analysis *Statistical Analysis of Medical Data Using SAS* (London: Chapman and Hall/CRC)

[62] Sanghani P, Ang B T, King N K K and Ren H 2018 Overall survival prediction in glioblastoma multiforme patients from volumetric, shape and texture features using machine learning *Surg. Oncol.* **27** 709–14

[63] Yang Z, Hu Z, Ji H, Lafata K, Floyd S, Yin F F and Wang C 2022 A neural ordinary differential equation model for visualizing deep neural network behaviors in multi-parametric MRI based glioma segmentation (arXiv:2203.00628)

[64] Li Y, Kim M M, Wahl D R, Lawrence T S, Parmar H and Cao Y 2021 Survival prediction analysis in glioblastoma with diffusion kurtosis imaging *Front. Oncol.* **11** 690036

[65] Horská A and Barker P B 2010 Imaging of brain tumors: MR spectroscopy and metabolic imaging *Neuroimaging Clin.* **20** 293–310

[66] Law M, Yang S, Wang H, Babb J S, Johnson G, Cha S, Knopp E A and Zagzag D 2003 Glioma grading: sensitivity, specificity and predictive values of perfusion MR imaging and proton MR spectroscopic imaging compared with conventional MR imaging *Am. J. Neuroradiol.* **24** 1989–98