



# BRAIN TUMOR DETECTION USING U-NET++ SEGMENTATION TECHNIQUE

Jayashree Shedbalkar<sup>1</sup>, Dr. K. Prabhushetty<sup>2</sup> Prof. A S Inchal<sup>3</sup>

<sup>1</sup> Department of Computer Science and Engineering, KLS VEDIT, Haliyal, VTU Belagavi, KARNATAKA, INDIA

<sup>2</sup> Department of Electronics Communication and Engineering, VTU Belagavi, KARNATAKA, INDIA

<sup>3</sup> Department of Computer Science and Engineering, KLS VEDIT, Haliyal, VTU Belagavi, KARNATAKA, INDIA

## ABSTRACT

Segmenting Brain Tumor images is crucial for computer-assisted diagnosis. The key to effective segmentation is for the model to be able to both see the overall picture and the minute details, or to learn image characteristics that contain a lot of contexts while maintaining high spatial resolutions. The most popular techniques, U-Net and its variations, extract and fuse multi-scale information in order to reach this aim. The fused features performance is nonetheless constrained by their tiny effective receptive fields and emphasis on local visual signals. In this work, we use a variety of machine learning techniques to forecast the survival rate. To conduct segmentation, we use a 3D UNet++ architecture and combine channel and spatial attention with the decoder network. To forecast the length of each patient's survival, we extract certain unique radiomic parameters based on the geometry, position, and shape of the segmented tumor and integrate them with clinical data. To demonstrate the impact of each attribute on the prediction of overall survival (OS), we also conduct comprehensive studies. According to the experimental findings, the most important factors to determine the OS are clinical characteristics like age and radiomics properties like the histogram, location, and shape of the necrosis area.

we offer Segtran, a different segmentation framework built on transformers and UNet++, which even at high feature resolutions have an infinite effective receptive field. Segtran's central component is a new squeeze-and-expansion UNet++, in which an expansion block learns a variety of representations while a squeezed attention block regulates the self-attention of transformers. We also provide a brand-new positional encoding approach for transformers that imposes an image continuity inductive bias.

## Keywords:

*Brain tumor segmentation, Glioma, Regression, Attention, Rate of Survival*

## 1. INTRODUCTION

The most prevalent and most fatal type of brain tumours are gliomas, which arise from glial cells. Nearly 190,000 instances of gliomas are reported on average each year worldwide [5]. The typical patient survival period for glioma patients is still about 12 months [7], and after 24 months following surgical resection, roughly 90% of patients had passed away [16]. For survival prediction and treatment planning, early identification, automated delineation, and volume estimate are essential responsibilities. However, because of their wide variety in shape, location, and appearance, gliomas are sometimes challenging to detect and define using traditional manual segmentation. The segmentation of tumor tissue must also be manually annotated, which takes time and effort and requires careful human expert supervision. The diagnosis and treatment will be considerably more accurate and swifter with the use of automatic segmentation and survival rate prediction models.

U-Net++ [Ronneberger et al., 2019], which was introduced, has demonstrated exceptional performance in a variety of Brain Tumor Medical picture segmentation tasks. A U-Net++ is made up of an encoder and a decoder. The encoder creates coarse contextual features that concentrate on contextual patterns by gradually downsampling the data, while the decoder gradually upsamples the contextual features and fuses them with fine-grained local visual characteristics. The RF of U-Net++ is expanded by the incorporation of various scale elements, which accounts for its successful performance. However, the influence of distant pixels soon fades as the convolutional layers are deepened. Because of this, a U-effective Net's RF is substantially lower than its theoretical RF. The effective RFs of a regular U-Net++ and DeepLabV3+ are only around 90 pixels, as illustrated in Fig. 2. This suggests that they struggle to represent greater context and base their judgments mostly on isolated, tiny patches. The heights/widths of the ROIs, however, are frequently higher than 200 pixels in many workloads, much over their actual RFs. U-Net++ and other models may be fooled by regional visual signals and commit segmentation mistakes if they lack the ability to "see the wider picture".

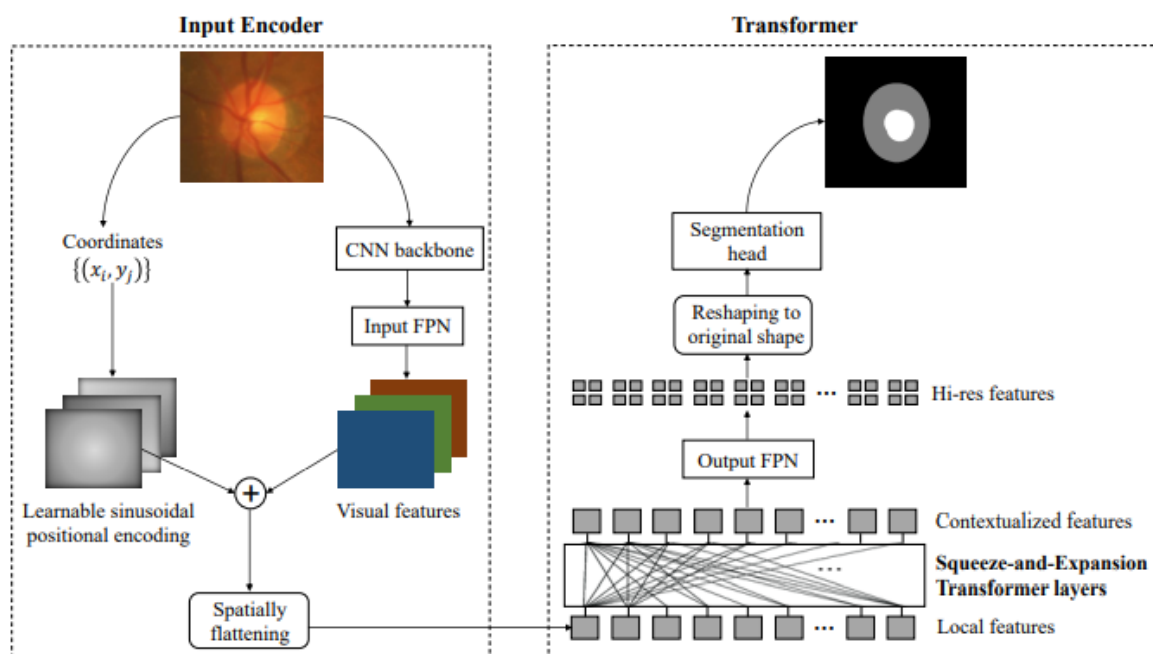


Figure 1: Segtran Brain Tumor Architecture

Through the use of a CNN backbone, it extracts visual features, merges them with positional encodings of the pixel locations, and flattens the resulting set of local feature vectors. A few layers of Squeeze-and-Expansion transformers contextualise the local characteristics. An input FPN and an output FPN upsample the features before and after the transformers to boost spatial resolution.

Transformers are becoming more and more common in computer vision tasks [Vaswani et al., 2019]. A transformer mixes the characteristics of all the input units, calculates the pairwise interactions between them, and then produces contextualized features. With the exception of its limitless effective receptive field, which is adept at catching long-range correlations, the contextualization provided by a transformer is comparable to the upsampling path in a U-Net. Transformers are thus an obvious choice for picture segmentation. In this study, we introduce Segtran, an alternative transformer-based segmentation architecture.

Simple segmentation with transformer integration only produces mediocre performance. Transformers might be modified in a number of ways to better serve picture applications because they were first developed for Natural Language Processing (NLP) activities. In order to do this, we suggest a unique transformer design called the Squeeze-and-Expansion Transformer. In this design, an expansion block learns a variety of representations while a compressed attention block helps regularise the enormous attention matrix. A learnable sinusoidal positional encoding that imposes a continuous inductive bias for the transformer is also something we suggest. The results of experiments show that they enhance segmentation performance.

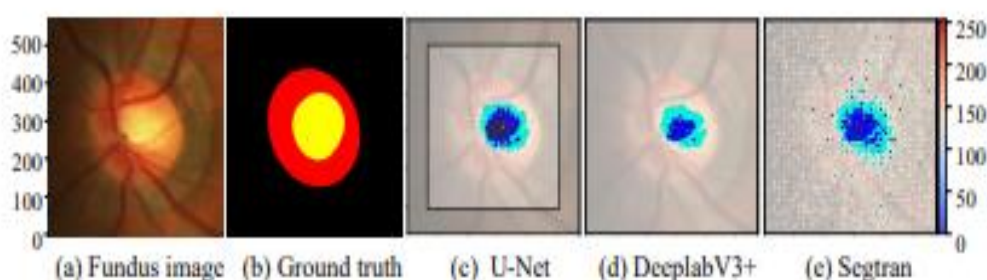


Figure.2 : Three model's three effective receptive fields are shown, with no discernible gradients in the blue blobs and light-colored dots. Back propagation of gradients occurs from the image's focal point. Segtran's picture contains barely perceptible gradients scattered across it (light-colored dots). Concentrated gradients are seen in U-Net++ and DeepLabV3+. Image input size: 576X576.

We assessed Segtran on two 2D medical picture segmentation tasks: polyp segmentation in colonoscopy images and optic disc/cup segmentation in fundus images from the REFUGE'20 competition. We also tested it on a 3D image segmentation problem, namely the segmentation of brain tumors on MRI data from the BraTS'19 challenge. According to Chen et al 2018 's research, Segtran consistently outperforms U-Net, U-Net++, UNet3+, PraNet, and U-Net as well as DeepLabV3+.

## 2. METHOD

DETR [Carion et al., 2020] served as a major inspiration for our work. DETR learns a series of object queries to extract the locations and classes of objects in an image using transformer layers to provide contextualised features that represent objects. While DETR is investigated for panoptic segmentation as well [Kirillov et al., 2019], its two-stage technique is not appropriate for segmenting brain tumor images.

When it comes to Brain Tumor picture segmentation, Cell-DETR [Prangemeier et al., 2020], a follow-up to DETR, also uses a transformer. However, Cell-architecture DETR's is essentially a simplified DETR and does not include any innovative elements like our Squeeze-and-Expansion transformer. Most recently, SETR and TransU-Net were published simultaneously with or after our research was submitted (Zheng et al., 2021; Chen et al., 2021).

Both of them use a Vision Transformer (ViT) as the encoder to extract picture features, which already contain global contextual information [Dosovitskiy et al., 2021]. The segmentation mask is created using a few convolutional layers as the decoder. On top of the local image features that were collected from a CNN backbone in Segtran, the transformer layers provide global context, and a Feature Pyramid Network (FPN) creates the segmentation mask.

CNNs are extended with positional encoding channels in [Murase et al., 2020], and they are assessed on segmentation tasks. Results were inconsistent. On the other hand, we demonstrated through an ablation research that positional encodings do, in fact, assist Segtran in doing segmentation to a certain extent.

More downsampling layers can be used to increase U-Nets' receptive fields. The danger of overfitting is increased as a result of the increased number of factors. Increasing the stride widths of the convolutions in the downsampling process is another approach to expand receptive fields. But by doing so, feature map spatial accuracy is sacrificed, which is frequently bad for segmentation [Liu and Guo, 2020].

**2.1 SQUEEZE AND EXPANSION TRANSFORMER**

Self-Attention, which may be thought of as computing an affinity matrix between various units and utilizing it to collect features, is the fundamental idea of a transformer.

$$Att\_weight(X, Y) = f(K(X), Q(X) \in R^{N \times N} \text{ -----(1)}$$

$$Attention(X) = Att\_weight(X, Y) \cdot V(X) \text{ -----(2)}$$

$$X_{out} = FNN(Attention(X)) \text{ -----(3)}$$

key, query, and value projections, respectively, are represented by K, Q, and V. Softmax is the product following the dot. Its I jth element specifies how much the characteristics of unit j contribute to the fused (contextualised) features of unit i. Att weight(X, X) is the pairwise attention matrix between input units. A feedforward network called FFN is utilised to further manipulate fused features.

In order to capture various kinds of linkages between input units, the basic transformer described above is expanded to a multi-head attention (MHA) [Vaswani et al., 2017; Voita et al., 2019]. The individual attention weights and output features (C/Nh-dimensional) computed by each of the Nh heads are then concatenated along the channel dimension to produce the C-dimensional features. Different heads only function in certain feature subspaces. We contend that by making four improvements, transformers may be made to be more suitable for images:

The projected input features in Eq. (2) are linearly combined to create the intermediate features Attention(X), with the attention matrix defining the combination coefficients. The attention matrix is naturally susceptible to noise and overfitting because it is so large: N N, with N generally > 1000. It could be beneficial to reduce the attention matrix to lower rank matrices.

Traditional transformers contain monomorphic output features, which may not have enough capacity to adequately characterise data changes because there is only one set of feature transformations (the multi-head transformer also has one set after concatenation). A combination of k transformers can better capture data changes than a single k transformer, just as a mixture of Gaussians nearly always does.

Traditional transformers can recognise asymmetric connections between tokens in natural language since the key and query projections are separately learnt. However, there are frequently symmetrical connections between picture units, such as whether two pixels are members of the same segmentation class.

A pixel's locality and semantic continuity are quite strong. Such an inductive bias is not entirely imposed by the two widely used positional encoding systems [Carion et al., 2020; Dosovitskiy et al., 2021]. A positional encoding improvement may introduce this bias.

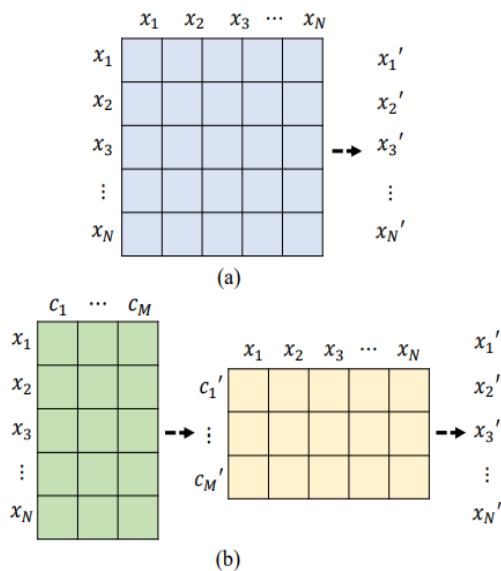


Figure 3 : (a) Squeezed Attention Block vs. (b) Full Self-Attention (N N) (SAB). In SAB, initial input units  $x_1, \dots, x_N$  interact with a codebook  $c_1, \dots, c_M$  to produce projected codebook features  $c_{0,1}, \dots, c_{0,M}$ , which then interact again with the input  $x_1, \dots, x_N$ . The two attention matrices are, respectively,  $N \times M$  and  $M \times N$ .

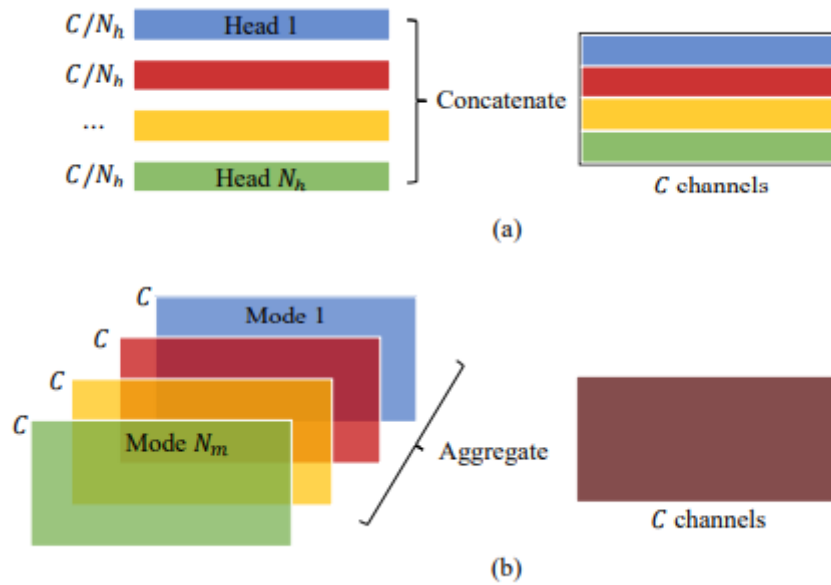


Figure 4 : Expanded attention block vs. Multi-head attention (MHA) (EAB). Each head in MHA generates a distinct feature subset. EAB, on the other hand, combines  $N_m$  sets of full characteristics from  $N_m$  modes and outputs them.

All four areas are where the Squeeze-and-Expansion Transformer seeks to enhance. The Squeezed Attention Block compresses the attention matrices to  $N \times M$  and computes attention between the input and  $M$  inducing locations [Lee et al., 2019]. The Expanded Attention Block uses  $N_m$  modes, or "experts," in a mixture-of-experts paradigm. To better reflect the symmetric interactions between picture units, the query projections and key projections are coupled in both blocks to make the attention symmetric. The model also benefits from a Learnable Sinusoidal Positional Encoding to capture spatial correlations.

**2.2 SEGTRAN ARCHITECTURE**

Segmentation, a context-dependent pixel-by-pixel classification job, must choose between localization accuracy and greater context (lower resolution) (higher resolution). Segtran partially resolves this dilemma without sacrificing spatial resolutions by doing pairwise feature contextualization. Segtran is made up of five primary parts (Fig. 1): Feature extraction from images is accomplished through a CNN backbone, input/output feature pyramids, learnable sinusoidal positional encoding, squeeze-and-expansion transformer layers, a segmentation head, and learnable sinusoidal positional encoding.

**2.2.1 : CNN Backbone**

In order to extract features maps with rich semantics, we use a pre-trained CNN backbone. Assume the input picture is  $X_0 \in R^{H_0}$ , where  $D_0$  is the number of colour channels in a 2D image and is either 1 or 3. The number of slices in the depth dimension of a 3D picture is  $D_0 = 1$  or 3. The retrieved features for 2D and 3D pictures are  $CNN(X_0) X_0 \in R^{H_0}$ , respectively.

Usually, ResNet-101 or EfficientNet-D4 serve as the foundation for 2D pictures. We reduce the stride of the first convolution from 2 to 1 to improve spatial resolution.  $H, W$  then become  $H_0/16$  and  $W_0/16$ . 3D backbones like I3D [Carreira and Zisserman, 2017] might be used to 3D pictures.

**2.2.2 : TRANSFORMER LAYER**

Each unit's visual characteristics and positional encodings are combined before being given into the transformer, as follows:  $X_{spatial} = X_{visual} + pos(coordinates(X))$ . A 1-D sequence,  $X_0 \in R^{Nu \times C}$ , is created by flattening spatial across spatial dimensions, where  $Nu$  is the total number of picture units, or the points in the feature maps.

A few stacked transformer layers make up the transformer. Each layer receives a set of contextualised features  $X$  as input, computes the pairwise interactions between the input units, and returns the same set of contextualised features  $X$ . Our unique Squeeze-andExpansion Transformer design is employed for the transformer layers.

**2.2.3: PYRAMIDS OF FEATURES AND SEGMENTATION HEAD**

The input features to transformers are typically high-level characteristics from the backbone, even if the spatial resolution of features does not decrease after passing through the transformer layers for richer semantics. However, they have a poor spatial resolution. So, using an input Feature Pyramid Network (FPN) and an output FPN that upsample the feature maps at the transformer's input end and output end, respectively, we boost their spatial resolution.

Let's suppose the EfficientNet is the backbone to preserve generality. Multi-scale feature maps are frequently extracted from the network during stages 3, 4, 6, and 9. The associated feature maps will be referred to as  $f_1, f_2, f_3$ , and  $f_4$ , correspondingly.

As mentioned below,  $f(X_0) = f_4$  is  $1/16$  is too coarse for precise segmentation of the source picture. So, using an input FPN, we upsample it to create upsampled feature maps  $f_{34}$ :

$$f_{34} = \text{upsample}_{x2}(f_4) + \text{Conv}_{34}(f_3)$$

where  $upsample_{x2}$  is bilinear interpolation and  $Conv_{34}(f_3)$  convolution that aligns the channels of  $Conv_{34}(f_3)$ .

$f_{34}$  is utilised as the input features to the transformer layers and represents 1/8 of the original picture. The output feature maps are 1/8 of the input picture since the transformer layers maintain the spatial resolutions from input to output feature maps. However, segmentation cannot be done with this spatial resolution. In order to upsample the feature maps by a factor of 4 (i.e., 50% of the original pictures), we adopt an output FPN:

$$f_{12} = upsample_{x2}(f_2) + Conv_{12}(f_2)$$

$$f_{1234} = upsample_{x2}(f_{34}) + Conv_{34}(f_{12})$$

When  $f_1$  to  $f_2$  and  $f_2$  to  $f_4$  channels, respectively, are aligned by 1X1 convolutional layers conv12 and conv24.

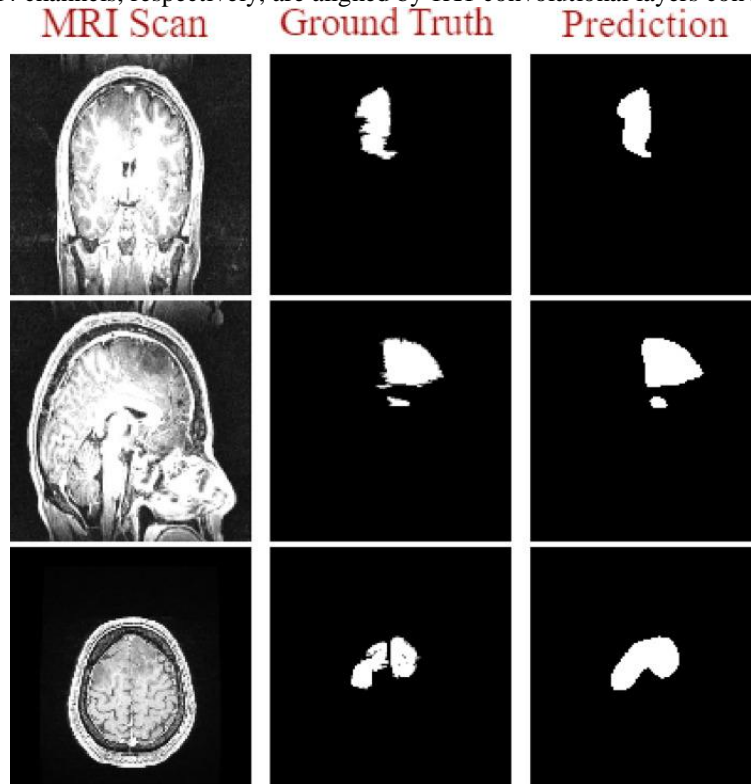


Figure 4 : Tumor segmented ground truth and prediction

### 2.3: SEGMENTATION

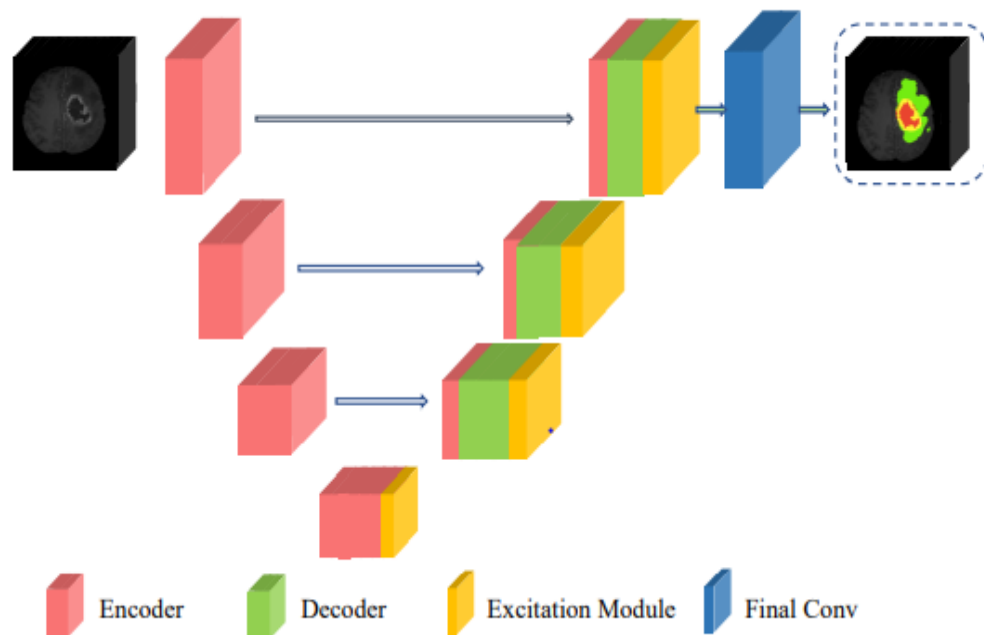


Figure 5 : Our suggested segmentation design, the 3D Attention UNet, combines a spatial attention mechanism with a sequential channel.

We adapt the UNet++ [20] design to 3D, combine the 3D attention module with the decoder blocks, and use it. To further improve segmentation prediction, we also provide a 3D attention model with decoder blocks [8]. The attention module that we

suggest combines channel and spatial attention with skip connection. However, combining intriguing features in simultaneously might lead to inconsistent feature learning. As shown in Fig. 2, integrating skip connection decreases the network's redundancy and sparsity. In Fig. 5, the total architecture is depicted.

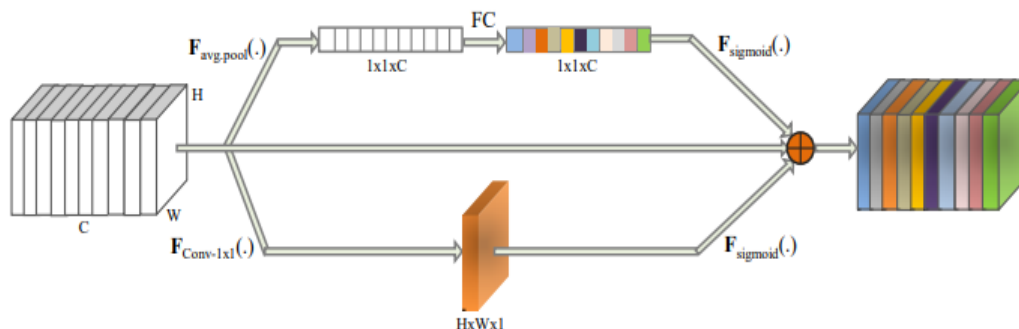


Figure 6 : Visualization of the 3D spatial and attentional channels with skip connectivity

3D Skip Attention Unit, Through its feature hierarchy, spatial and channel attention improves the quality of encoding. In order to produce 3D spatial and channel attention, we thus offer 3D attention units that take use of 3D inter-spatial and inter-channel feature correlations (as illustrated in Fig. 6). We first carry out a 1x1xC convolution to combine all spatial feature correlations into the HxWx1 dimension in order to produce the 3D attention map. In order to obtain the 1x1xC channel correlation, we execute average pooling in parallel and input it to the neural network.

The 3D attention map that has been encoded has detailed spatial and channel attention. We also combine skip-connection to lessen the singularity and sparsity brought on by these parallel excitations. Additionally, by using skip connections, the learning becomes more general while improving segmentation prediction.

### 2. 4: SURVIVAL PREDICTION

Feature Extraction, As in our earlier research [10], characteristics that reveal the geometry and fractal structure of the tumour have a significant effect on the number of days of survival. The set of characteristics utilised in [10] together yields the overall survival (OS) prediction job for BraTS 2018 validation challenge with the highest degree of accuracy. However, the strategy didn't work during the BraTS 2019 test phase since the data on the regression model were overfit. Therefore, in this work improving learning approaches, the same mix of characteristics is applied. The lengths and coordinates of the first, second, and third axes are retrieved as geometrical characteristics. For necrosis, the tumour core, and the whole tumour, centroid coordinates, eigenvalues, meridional and equatorial eccentricity, fractal dimensions, and histogram properties including entropy, skewness, and kurtosis are also retrieved. For the purpose of avoiding magnitude disparities, all characteristics are normalised to the 0–1 range.

Feature Selection, The most important features for predicting survival must be fed into the regression model in order to maximise prediction accuracy. So, for feature ranking, we investigate recursive feature elimination (RFE). Getting the most important characteristics is the method's main goal. The number of characteristics is gradually raised in order to determine the ideal number, which mostly pertains to the overall survival (OS) prediction job.

UNet ++ Model, To forecast the overall survival, we use the cutting-edge UNet model [6] on the chosen characteristics (OS). To get the best-performing model, we adjust the hyperparameters, such as the maximum tree depth, learning rate, verbosity level, and L1 and L2 regularisation terms on weights. The use of regularisation terms is advantageous in regression tasks because L1 and L2 terms regulate sparsity and over-fitting. We also use a number of additional machine learning techniques that are frequently employed for regression challenges. For instance, random forest (RF) [15], support vector machine (SVM) [22], and multi-layer perceptron (MLP) [15].

### 3. RESULTS AND DISCUSSION

We enter our model prediction into the BraTS 2019 site and retrieve many measurement metrics, including Dice, Hausdorff, Sensitivity, and Specificity, to assess the accuracy of our model prediction. Table 1 provides an example of the BraTS 2019 validation set's performances. Fig. 3 provides an illustration of the validation set prediction visualisation. Fig. 4 depicts the performance graph of the original 3D UNet and our suggested 3D attention UNet. It is evident that for all areas, including ET, WT, and TC, 3D attention UNet surpasses the original model.

Table 2 displays the quantitative outcomes for the BraTS 2019 test set. We may deduce our model's forecast for the BraTS 2019 testing dataset from Fig. 7.

Table 1 : BraTS 2019's validation set is evaluated using the Dice, Hausdorff, Sensitivity, metrics for the segmentation job.

	Dice			Hausdorff			Sensitivity		
	ET	WT	CT	ET	WT	CT	ET	WT	CT
Mean	0.7132	0.9223	0.8023	8.02	7.26	9.63	0.7632	0.9125	0.8236
Std	0.3225	0.0821	0.1963	14.23	12.36	14.23	0.2986	0.089	0.1965
Median	0.8235	0.9165	0.7785	2.36	4.32	5.23	0.8691	0.9265	0.9015

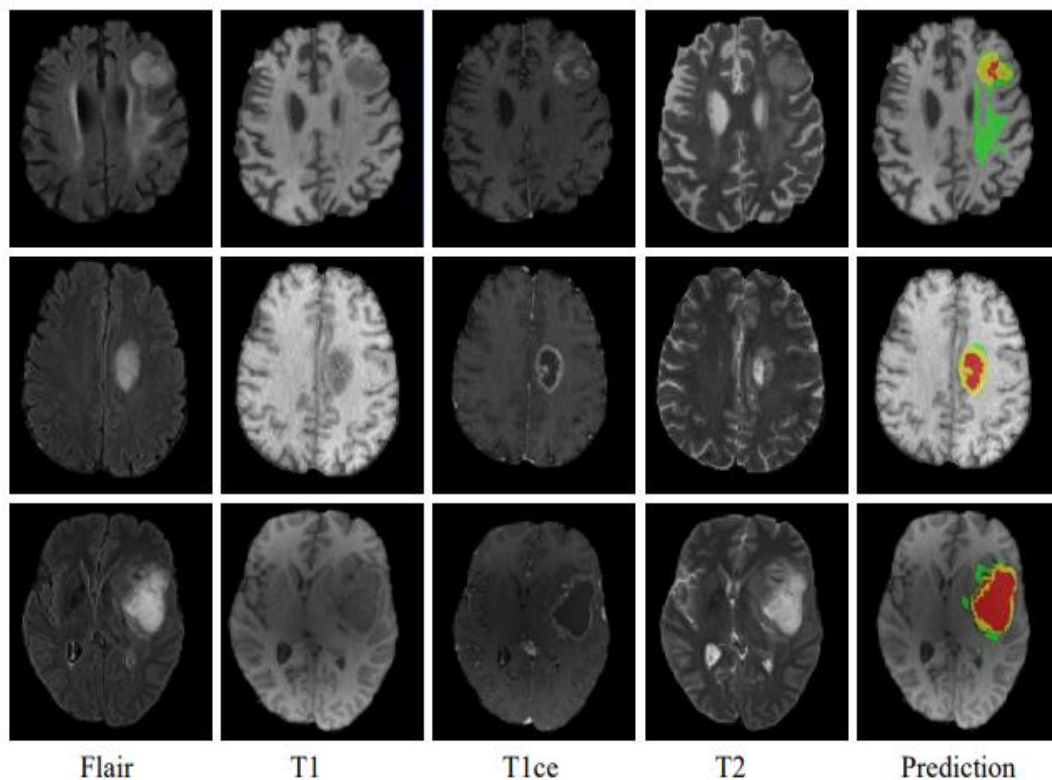


Figure. 7: Using the Ground-Truth and Predicted segmentation of tumour sub-regions for the BraTS 2019 testing dataset, the Flair, T1, T1ce, and T2 modalities of the brain tumour were depicted. Red denotes necrosis, yellow an enhanced tumour, and green denotes edoema, according to the colour of the annotation.

Table 2 : Evaluation of the BraTS 2019 testing set for the segmentation challenge using dice and Hausdorff metrics

	Dice			Hausdorf		
	ET	WT	CT	ET	WT	CT
Mean	0.7896	<b>0.8796</b>	0.7896	4.5621	8.3697	7.8954
StdDev	0.2236	0.1563	0.2956	6.2358	13.2540	12.3215
Median	0.8456	0.9215	0.9023	3.2154	4.2154	4.2514

Table 3 : UNet++, and Random Forest (RF) performance comparison on validation set for overall survival prediction. The terms MSE and stdSE refer for the projected survival days' mean square error and standard deviation.

Method	Accuracy	MSE	MedianSE	stdSE
UNet++	74.53%	120025.118	49555.12	318289.21
Random Forest	69.32%	110283.239	50647.00	246936.02

### 3.1: SURVIVAL PREDICTION

Our study uses a number of cutting-edge regression methods to assess the survival rate. Although there are 125 examples in the validation set, only 29 anonymous cases are used in the BraTS 2019 assessment site to verify the model. In order to assess the regression model on the training dataset, we performed 4-fold cross-validation. The performance comparison of all the models is shown in Table 3. While MLP obtains the lowest MSE, Random Forest exceeds all other regression models with the greatest accuracy. To analyse the validation and test set while taking performance into account, we choose UNet++. The performance of the UNet++ OS on the BraTS 2019 test and validation dataset is shown in Table 4.

### 3.2 DISCUSSION

According to the findings, our 3D attention UNet generates more accurate results than the original 3D UNet. Particularly, the prediction of the tumour core, which is a crucial area to determine malignant prognosis, increases in our approach (as shown in Fig. 4). We use 4 different regression models to estimate the OS, with Random Forest doing better in terms of accuracy. We choose the top 14 characteristics and train the models to create an effective model.

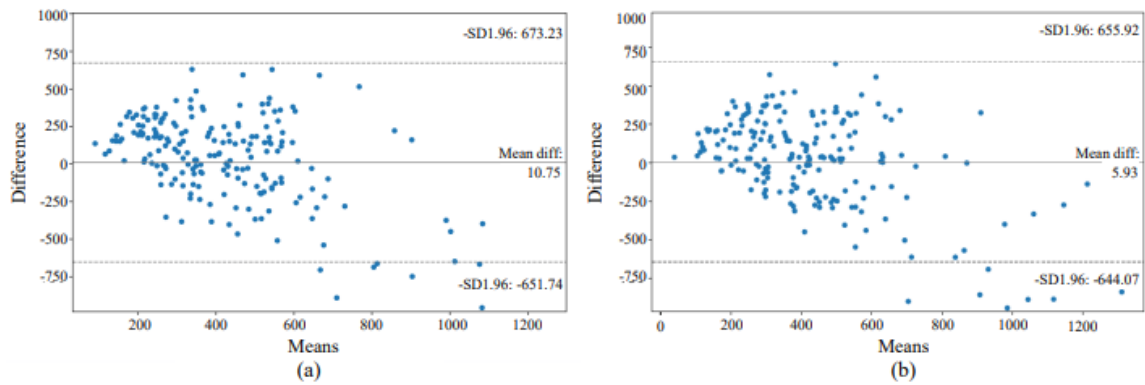


Figure 8 : Bland Altman plot was produced for all extracted features using the training cross-validation findings of the overall survival prediction model (a). A mean difference of 10.75 days results from this. (b) The Bland Altman plot for the 14 characteristics that were chosen. A mean difference of 5.95 days results from this.

The distribution of regression output for all extracted features and 14 chosen features is shown as a Bland Altman plot in Fig. 6 (a and b). When comparing the selected features to all features, the average gap between the actual survival rate and the anticipated survival rate is over half (5.93 days).

The selected characteristics' significance for the performance of the model is shown in Fig. 7. An way to explaining the results of tree ensemble methods like Random Forest is called SHAP (SHapley Additive exPlanations) analysis, which is based on game theory. Red denotes high feature values, whereas blue denotes low feature values. The 14 characteristics chosen for our tests are displayed on the plot's y-axis. We may conclude that age contributes most significantly to model performance. Additionally, the overall tumour volume, eigenvalue, histogram of necrosis, and second axis length of the tumour voxel are some of the important parameters that helped predict OS. The regression plot of the model's prediction and the ground truth is displayed in Fig. 8.

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [14], [15]. The discussion can be made in several sub-sections.

#### 4. CONCLUSION

We describe a segmentation and survival prediction algorithm in this study for MRI-based automated brain tumour prognosis. We use UNet and combine the 3D attention technique to provide a brand-new approach to capturing the key elements in model learning. In order to estimate the survival days using the regression model, we additionally extract a variety of innovative geometric and morphological variables. We note that the most important factors in estimating the prognosis of gliomas are the location, shape, and size of the necrotic region.



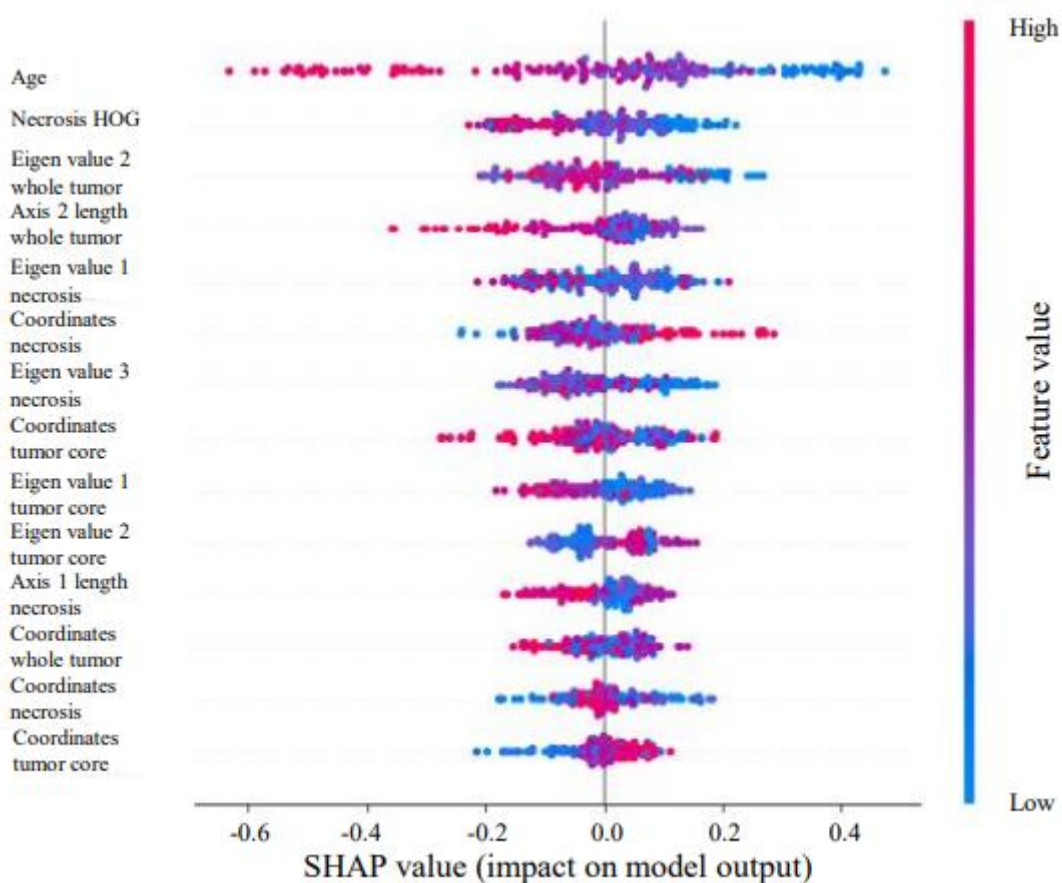




Figure. 9 : Effect of the characteristics on the model's output. To assess the importance of the features in model prediction, the colours red and blue reflect the high and low feature values, respectively

## REFERENCES

- [1] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. *The Cancer Imaging Archive* 286(2017)
- [2] 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 mricollections with expert segmentation labels and radiomic features. *Scientific data* 4, 170117 (2017)
- [3] Castells, X., Garc'ia-G'omez, J.M., Navarro, A., Acebes, J.J., Godino, O., Boluda, S., Barcel'o, A., Robles, M., Ari'no, J., Ar'us, C.: Automated brain tumor biopsy prediction using single-labeling cDNA microarrays-based gene expression profiling. *Diagnostic Molecular Pathology* 18(4), 206–218 (2009)
- [4] Islam Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 785–794. ACM (2016)
- [5] , M., Atputharuban, D.A., Ramesh, R., Ren, H.: Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robotics and Automation Letters* 4(2), 2188–2195 (2019)
- [6] 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 (2015)
- [7] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
- [8] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 234–241. Springer (2015)
- [9] Ruck, D.W., Rogers, S.K., Kabrisky, M., Oxley, M.E., Suter, B.W.: The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Transactions on Neural Networks* 1(4), 296–298 (1990) Chen et al., 2021]
- [10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306,
- [11] Esser et al., 2020] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. arxiv:2012.09841, 2020. [Fan et al., 2020]
- [12] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In MICCAI, 2020.
- [13] [Huang et al., 2020] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. Chen, and J. Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In ICASSP, 2020.
- [14] Isensee et al., 2021] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a selfconfiguring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18, 2021
- [15] Sun'ao Liu and Xiaonan Guo. Improving brain tumor segmentation with multi-direction fusion and fine class prediction. In BrainLes workshop, MICCAI, 2020
- [16] Rito Murase, Masanori Suganuma, and Takayuki Okatani. How Can CNNs Use Image Position for Segmentation? arXiv:2005.03463, 2020
- [17] Tim Prangemeier, Christoph Reich, and Heinz Koepl. Attention-based transformers for instance segmentation of cells in microstructures. In *IEEE International Conference on Bioinformatics and Biomedicine*, 2020

- [18] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnet: Split-attention networks. arXiv:2004.08955, 2020]
- [19] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In CVPR, 2021.
- [20] Kutlu H, Avci E. A Novel Method for Classifying Liver and Brain Tumors Using Convolutional Neural Networks, Discrete Wavelet Transform and Long Short-Term Memory Networks. Sensors (Basel). 2019 Apr 28;19(9):1992. doi: 10.3390/s19091992. PMID: 31035406; PMCID: PMC6540219
- [21] Ker J, Bai Y, Lee HY, Rao J, Wang L. Automated brain histology classification using machine learning. J Clin Neurosci. 2019 Aug;66:239-245. doi: 10.1016/j.jocn.2019.05.019. Epub 2019 May 31. PMID: 31155342.
- [22] Hussain Z, Gimenez F, Yi D, Rubin D. Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. AMIA Annu Symp Proc. 2018 Apr 16;2017:979-984. PMID: 29854165; PMCID: PMC5977656.
- [23] Kammerlander C, Neuerburg C, Verlaan JJ, Schmoelz W, Miclau T, Larsson S. The use of augmentation techniques in osteoporotic fracture fixation. Injury. 2016 Jun;47 Suppl 2:S36-43. doi: 10.1016/S0020-1383(16)47007-5. PMID: 27338226.
- [24] Zeng S, Zhang B, Gou J, Xu Y. Regularization on Augmented Data to Diversify Sparse Representation for Robust Image Classification. IEEE Trans Cybern. 2020 Oct 21;PP. doi: 10.1109/TCYB.2020.3025757. Epub ahead of print. PMID: 33085628.
- [25] Nalepa J, Marcinkiewicz M, Kawulok M. Data Augmentation for Brain-Tumor Segmentation: A Review. Front Comput Neurosci. 2019 Dec 11;13:83. doi: 10.3389/fncom.2019.00083. PMID: 31920608; PMCID: PMC6917660.
- [26] Porter ND, Verdery AM, Gaddis SM. Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. PLoS One. 2020 Jun 10;15(6):e0233154. doi: 10.1371/journal.pone.0233154. PMID: 32520948; PMCID: PMC7286483.

## BIOGRAPHIES OF AUTHORS

	<p>JAYASHREE SHEDBALKAR has received her Bachelor's degree in Computer Science and Engineering from Visvesvaraya technological University, Belagavi, Karnataka Masters in Computer Network and Engineering from from VTU Belgaum. Currently she is an Assistant Professor CSE dept. at KLS VEDIT Haliyal, her research interests include Image Processing and Artificial Intelligence</p>
	<p>Dr. K. Prabhushetty received his PhD in Electronics from Shivaji University Kolhapur, India. He is having more than three decades of experience in the field of engineering and technology. His research area is Medical Image Processing. He is guiding 7 PhD scholars. Currently he is working as professor in department of Electronics &amp; Communication Engineering affiliated to VTU Belgaum</p>