



Article

An Interpretable Stacked Deep Learning Model for Diagnosis of Brain Tumor with Transparent Learning Dynamics

K. Kaivalya ¹, N. Thirupathi Rao ¹, Aditya Pal ^{2,*}, Hari Mohan Rai ^{3,*}  and B. Omkar Lakshmi Jagan ⁴

¹ Department of Computer Science & Engineering, Vignan's Institute of Information Technology (A), Visakhapatnam 530049, India; 21131a0585@vignaniit.edu.in (K.K.); nakkathiru@vignaniit.edu.in (N.T.R.)

² Department of Convergent Environmental Science, College of Life Science and Biotechnology, Dongguk University, Seoul 04620, Republic of Korea

³ School of Engineering and Digital Sciences, Department of Computer Science, Nazarbayev University, Astana 010000, Kazakhstan

⁴ Department of Electrical & Electronics Engineering, Vignan's Institute of Information Technology (A), Visakhapatnam 530049, India; omkarjagan@vignaniit.edu.in

* Correspondence: aditya@dgu.ac.kr (A.P.); hari.ra@nu.edu.kz (H.M.R.)

Abstract

The diagnosis and treatment planning for brain tumors remain a complex task in medical imaging, largely due to the intricate structure of such abnormalities. This study introduces an interpretable stacked deep learning framework consisting of three sequential stages: (i) tumor segmentation, (ii) feature extraction, and (iii) tumor classification. The segmentation stage introduces a three-parameter lambda distribution (TPLD), a symmetric special case of generalized lambda distribution (GLD), used as a statistical intensity prior that is fused into the gating signal of an Attention U-Net for enhancing boundary delineation. The segmented outputs are processed using InceptionV3 for deep feature extraction and followed by a convolutional neural network (CNN) classifier. We evaluated the proposed model on the BRISC 2025 dataset, consisting of T1 weighted brain MRI images with pixel wise segmentation masks, which is validated by medical experts. The dataset consists of 3933 training images and 860 test images with ground truth masks, containing the four classes: meningioma, glioma, no tumor, and pituitary tumor. We utilized a region-of-interest-based training strategy to reduce the computational complexity and minimize overfitting. The data split followed the official image-level partition distributed with BRISC 2025; because patient identifiers are not released with the dataset, patient-level separation could not be independently verified, and this is acknowledged as a limitation. To ensure methodological transparency and clinical robustness, we systematically report the learning dynamics across 20, 60, and 100 training epochs at multiple decision thresholds (0.50, 0.60, 0.70), providing evidence of stable model convergence without overfitting. We also introduce a composite loss function by integrating cross-entropy, focal losses, and Dice to further boost performance. Experimental results demonstrate 97.8% classification accuracy on the test set, 92.4% Dice coefficient, and 85.9% IoU at the optimal threshold of 0.60. An ablation study further confirms the contribution of each loss component, supporting reproducibility and transparency in model evaluation. These findings confirm the practical utility and reliability of the proposed framework in the context of brain tumor segmentation and clinical diagnosis.



Academic Editors: Milan Toma and
Constantinos Pattichis

Received: 11 May 2026

Revised: 24 June 2026

Accepted: 30 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article
distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: brain tumor diagnosis; artificial intelligence; learning dynamics; clinical robustness; clinical transparency; interpretable stacked deep learning model; lambda distribution

1. Introduction

A brain tumor arises from the uncontrolled proliferation of the cells in or near brain tissue and is classified as primary (originating in the brain) or secondary/metastatic (spreading from elsewhere in the body) [1,2]. Tumors differ widely in type, size, growth rate, and location, and primary tumors are themselves divided into benign and malignant forms; benign tumors do not contain cancer cells but can still compress the surrounding structures as they enlarge, while malignant tumors invade and damage the healthy tissue more rapidly [2]. Diagnosis conventionally relies on CT and MRI imaging [3], with biopsy used to confirm tumor type and grade before the treatment, which typically combines surgery, radiation, and chemotherapy [4]. Prognosis varies sharply by tumor type: patients with glioblastoma multiforme have a median survival of only a few months, whereas benign meningiomas typically respond well to treatment [5]. Globally, around 250,000 new brain tumor cases are diagnosed each year, and brain tumors are the second most common childhood cancer after acute lymphoblastic leukemia [6].

This clinical burden, combined with the subjectivity and the workload of manual radiological review, motivates automated, AI-assisted diagnostic support, which is the primary focus of this study.

1.1. Motivation

Timely identification of neurological anomalies is essential for enhancing the patient prognosis and lowering fatality rates associated with brain cancers. However, the accuracy of the manual diagnosis by radiologists may be compromised due to factors such as fatigue, varying levels of expertise, and the subtle nature of early-stage tumors. Furthermore, traditional visual inspection of the medical images is not only time-intensive but also susceptible to human error and adds to healthcare costs. To overcome these challenges, we presented AI-assisted learning dynamics tools for the classification and segmentation of the brain tumor as a reliable solution. This study proposes an interpretable and stacked architecture designed to improve brain tumor diagnosis. The proposed framework consists of three core components: a TPLD-based Attention U-Net for the segmentation of the brain tumor, an Inception V3 model for feature extraction, and a CNN for classification of the brain tumor.

This work addresses several commonly observed challenges in medical AI research related to segmentation and detection, particularly the lack of transparency in the diagnostic process.

Many models reported in the literature achieve high benchmark accuracy while reporting only a single final-epoch performance number, without showing the trajectory of training and validation curves that would let a reader judge whether that number reflects a stable, generalizable optimization process or a fortunate single run. The depth and consistency of training-process reporting varies considerably across the literature: some studies report a learning curve or evaluate more than one operating threshold, but it remains comparatively uncommon for a single study to report training behavior across multiple epoch checkpoints and multiple decision thresholds together as a standard, built-in part of the evaluation protocol rather than as an incidental detail. This is symptomatic of a broader pattern in the medical deep-learning literature, where a model that reaches high accuracy through memorization, unstable optimization, or a fortuitously chosen hyperparameter setting on a single run can be difficult to distinguish, from the reported metrics alone, from a model whose learning process is genuinely stable. Without learning curves and multi-threshold evaluation reported together systematically, such problems can remain hidden from readers and, ultimately, from the clinicians who would need to trust the model.

To address this gap, our proposed framework reports training curves at 20, 60, and 100 epochs, evaluates performance across three decision thresholds (0.50, 0.60, 0.70), and

includes an ablation study that isolates the contribution of each loss term. We position this multi-epoch, multi-threshold reporting practice not merely as an additional experiment but as a methodological standard that we believe medical imaging AI research should adopt more broadly, since it is what allows a reader to visually confirm that optimization proceeded smoothly, that the train/validation gap remained reasonable, and that extending training did not degrade generalization.

Additionally, the proposed stacking strategy enables the visualization of the most significant regions within input images that contribute to the model's predictions, thereby offering interpretability alongside improved diagnostic performance.

1.2. Contributions

The contributions of the work are presented as follows:

- Presented a stacking framework with TPLD Attention U-Net, InceptionV3, and CNN for diagnosis of brain tumor—achieving 97.8% accuracy and 92.4% Dice on the BRISC 2025 dataset.
- Introduced a three-parameter lambda distribution directly into the segmentation network, replacing the heuristic intensity assumptions with a statistically principled model sharpening tumor boundary delineation while keeping the memory and compute costs tractable.
- Introduced a composite loss function (cross-entropy + Dice + focal) that achieves measurable performance gains of up to 2.0% Dice and 3.2% IoU over the cross-entropy-only baseline, without modifying the network architecture, thereby enabling straightforward integration into the other segmentation pipelines.
- Presented transparent training and evaluation analysis through learning curves at 20, 60, and 100 epochs, assessed across three decision thresholds, showing stable convergence and the absence of overfitting—offering the level of reliability and interpretability required for clinical and regulatory trust.
- Generated attention maps to show the specific tumor regions corresponding to the ground truth, making the model more transparent and allowing clinicians to better understand, verify, and trust its decision-making process.

2. Related Works

Table 1 summarizes several previous studies related to brain tumor segmentation and classification [7], highlighting the datasets used and their corresponding performance metrics. The studies listed span different datasets (BRATS variants, TCIA, BRISC, and others), different imaging modalities, and a mix of segmentation-only, classification-only, and combined tasks; reported accuracy values are therefore not directly comparable to one another or to the proposed model in a controlled, head-to-head sense and are included to situate the proposed work within the broader literature rather than as a strict benchmark. We restrict our own quantitative claims to results obtained on BRISC 2025 under the protocol described in Section 3.

A review of the selected literature points out that the limited research has investigated brain tumor detection via an integrated approach combining LDAU-Net with Inception V3 and a CNN-based classifier. Nevertheless, achieving improved performance necessitates the integration of a mathematical framework for enhancing segmentation precision and classification accuracy. To this end, a three-parameter Lambda distribution mathematical model has been integrated into the proposed model, improving the segmentation of brain tumor images for improved accuracy. Our proposed model, evaluated on the BRISC 2025 dataset, achieves 97.8% test accuracy. We note that Transformer-based segmentation backbones such as TransUNet [8] and Swin-Unet [9] represent an additional, increasingly used family of ap-

proaches for medical image segmentation [10]. For a controlled same-benchmark comparison, Section 4.3 presents classification performance of the proposed model alongside seven recent studies that also used BRISC 2025; a comparison against Transformer-based segmentation backbones under matched compute remains identified as future work in Section 6.

Table 1. Literature review on brain tumor segmentation and classification.

Ref	Author/Year	Model	Dataset	Tasks (Segmentation/Classification)	Evaluation Metrics
[4]	Pereira et al. (2016)	Deep CNN	BRATS 2013/2015	Segmentation	Dice (2013): 0.88 (WT), 0.83 (TC), 0.77 (ET)
[5]	Banerjee et al. (2019)	VolumeNet	TCGA (from TCIA)	Both	Test Acc: 97.19%; Val Acc: 98%
[6]	Amin et al. (2020)	SVM	Harvard, RIDER, Local	Both	Acc: 97.1%; AUC: 0.98; Sens: 91.9%; Spec: 98.0%
[11]	Isensee et al. (2021)	nnU-Net	BraTS 2020	Segmentation	Dice: 88.95 (WT), 85.06 (TC), 82.03 (ET)
[12]	Khairandish et al. (2022)	Hybrid CNN-SVM	BRATS 2015	Both	Overall Acc: 98.4959%
[13]	Xun et al. (2022)	RGA-Unet	TCIA	Segmentation	Dice: 95.904%; Loss: 0.228
[14]	Gupta et al. (2022)	InceptionResNetV2 + RFT	Kaggle (Navoneel/Bhuvaji)	Both	Detection Acc: 99%; Class Acc: 98%
[15]	Gammoudi et al. (2023)	BNet'S	BraTS2020	Segmentation	Dice: 0.869 (WT), 0.829 (ET), 0.854 (TC)
[16]	Xun et al. (2024)	ARGA-Unet	TCIA	Segmentation	Dice: 97.581%; Loss: 0.212
[17]	Khaliki et al. (2024)	VGG16	Kaggle (SARTAJ)	Classification	Acc: 98%; F-score: 97%; AUC: 99%; Recall: 98%
[18]	Aamir et al. (2024)	Optimized CNN	3 Kaggle dataset	Both	Acc: 97.18%; Pre: 97%; Rec: 97%; F1: 97%
[19]	Islam et al. (2026)	Xception	Merged (Figshare, SARTAJ, Br35H)	Classification	Test Acc: 98.71%; Val Loss: 0.0496
[20]	Anand et al. (2026)	CNN (4 blocks)	Kaggle (Sartaj Bhuvaji)	Classification	Acc: 98.5%; Pre: 96.2%; Rec: 89.7%; F1: 92.2%
[21]	Uniyal et al. (2026)	MobileNet	Merged (Figshare, SARTAJ, Br35H)	Both	Acc: 96.6%; Training Acc: 97.78%; Val Acc: 94.89%
[22]	Chatterjee et al. (2026)	GP-ShuffleUNet	BraTS 2020/Figshare	Both	Acc: 98.74% (tumor-only); F1: 0.93; Median Dice: 0.720
[23]	Ke et al. (2026)	MCACNN-SVM	Kaggle	Classification	Acc: 98.01%; Pre: 97.93%; Rec: 97.85%; F1: 97.88%
[24]	Srinivas et al. (2026)	4-conv-1-dense-1-dropout CNN	Crystal Clean MRI (Kaggle)	Classification	Acc: 95.86%; Pre: 95.8%; Rec: 95.4%; F1: 95.5%

3. Materials and Methods

This section outlines the materials and methods used for the proposed framework presented in this study.

3.1. Data Collection

The performance of the proposed framework was assessed using the BRISC 2025 dataset [1], a high quality expert-annotated MRI dataset curated for brain tumor classification and segmentation. The dataset provides T1 weighted brain MRI images with pixel-wise segmentation masks validated by physicians and radiologists. For the segmentation task, the dataset includes 3933 training images and 860 test images with corresponding tumor masks. The class distribution for the segmentation subset is as follows:

- **Training set (3933 Images):** glioma $\approx$ 902, meningioma $\approx$ 1045, pituitary $\approx$ 1146, no_tumor $\approx$ 840.
- **Test set (860 Images):** glioma $\approx$ 218, meningioma $\approx$ 263, pituitary $\approx$ 258, no_tumor $\approx$ 121.

All images are available in three anatomical panels (axial, coronal, sagittal), enabling the model to learn cross-plane features. The accurate and comprehensive expert annotation and balanced class distribution make BRISC 2025 an ideal benchmark for developing robust classification and segmentation models.

Data split and patient-level vs. image-level partitioning: The BRISC 2025 dataset was used according to the official image-level training and test split distributed with the dataset. In the publicly released version of the dataset, patient identifiers are not provided;

consequently, patient-level splitting could not be independently verified by the authors. To reduce the risk of information leakage, images from the official test set were not used during training or validation at any stage, and the validation set used for threshold and epoch selection (Section 3.1, Table 2) was drawn exclusively from the official training split. We acknowledge this as a limitation: patient-level splitting is the preferred protocol for clinical imaging studies whenever patient identifiers are available, because image-level splits can permit images from the same patient to appear in both training and evaluation sets, which may inflate reported performance relative to a strictly patient-disjoint evaluation. We flag this explicitly so that readers can weigh it when judging generalizability, and we recommend that future releases of BRISC, or follow-up studies with access to patient identifiers, re-evaluate the proposed pipeline under a verified patient-level split.

Table 2. Reproducibility summary: dataset split, preprocessing, augmentation, architecture, optimization, and hardware settings used for the proposed framework.

Setting	Value
Dataset	BRISC 2025
Task	Brain tumor segmentation and four-class classification
Classes	Glioma, meningioma, pituitary tumor, no tumor
Training images with masks	3500
Validation images	433 (sampled from the official training split)
Test images	860 (official test split, untouched during training/validation)
Dataset split	Official image-level split; patient-level split not verified (Section 3.1)
Input image size	224×224 (classification branch)/ 256×256 (segmentation branch)
Preprocessing	Resizing, skull stripping, contrast normalization, Z-score normalization
Data augmentation	Rotation $\pm 15^\circ$, horizontal flip, width/height shift up to 10%, zoom 0.9–1.1, brightness variation $\pm 10\%$ (training set only)
Segmentation network	TPLD-guided Attention U-Net (Section 3.6)
Feature extraction model	InceptionV3 (ImageNet-pretrained convolutional base, fine-tuned)
Classification model	CNN classifier on pooled InceptionV3 features
Activation functions	ReLU (hidden layers), Sigmoid (segmentation output), SoftMax (classification output)
Loss function	Composite loss: cross-entropy + Dice loss + focal loss (Equation (16))
Loss weights	$w_1 = 0.2$, $w_2 = 0.3$, $w_3 = 0.5$ (grid search; Section 4); renamed from the original manuscript's $\lambda_1, \lambda_2, \lambda_3$ notation to avoid collision with the TPLD distribution parameters of the same name (Section 3.4)
Optimizer	Adam
Initial learning rate	1×10^{-4} (Adam optimizer; we use a 0.001 base rate for the U-Net stage)
Learning-rate scheduling	Reduce learning rate on validation-loss plateau
Batch size	8
Epochs reported	20, 60, and 100
Weight initialization	He (mean-shift residual attention, MSRA-style) initialization
L2 regularization (weight decay)	0.00005
Early stopping	Applied based on validation-loss plateau
Decision thresholds evaluated	0.50, 0.60, and 0.70
Selected operating threshold	0.60
Random seed	42
Implementation language	Python 3.11
Deep learning library	Keras/TensorFlow
Hardware	Intel Core i9-13900K CPU (3.00 GHz), 64 GB DDR5 RAM, NVIDIA RTX 5090 48 GB GPU

3.2. Data Preprocessing

To ensure consistency and enhance model performance, each MRI image undergoes a series of pre-processing transformations. Let $I(a, b)$ denote the 2D MRI slice of dimension $H \times W$; all images are resized to a fixed spatial dimension $H' \times W'$ using bilinear interpolation. For a target pixel coordinate (a', b') , the corresponding (generally non-integer) coordinate in the original image, (x, y) , is given by

$$x = \frac{a'}{W'} \cdot (W - 1), \quad y = \frac{b'}{H'} \cdot (H - 1) \quad (1)$$

The interpolated intensity is then a weighted combination of the four nearest integer-coordinate pixels:

$$I'(a', b') = \sum_{i=0}^1 \sum_{j=0}^1 w_{ij} I(\lfloor x \rfloor + i, \lfloor y \rfloor + j), \quad (2)$$

where the four bilinear weights $w_{00}, w_{01}, w_{10}, w_{11}$ are computed from the fractional parts $dx = x - \lfloor x \rfloor$ and $dy = y - \lfloor y \rfloor$ (both in $[0, 1]$) as

$$\begin{aligned} w_{00} &= (1 - dx)(1 - dy), \\ w_{01} &= (1 - dx)dy, \\ w_{10} &= dx(1 - dy), \\ w_{11} &= dx dy, \end{aligned} \quad (3)$$

so that $w_{00} + w_{01} + w_{10} + w_{11} = 1$ for any $dx, dy \in [0, 1]$, and w_{ij} is the weight applied to the pixel located i columns and j rows from $(\lfloor x \rfloor, \lfloor y \rfloor)$.

To standardize the intensity ranges across different scans, we apply Z-score normalization, which transforms each pixel intensity $I'(a', b')$ as

$$\hat{I}(a', b') = \frac{I'(a', b') - \mu}{\sigma} \quad (4)$$

where μ and σ are the mean and standard deviation, respectively, of pixel intensities within the brain region. This ensures zero mean and unit variance, highlighting textural differences. Skull stripping is another critical preprocessing step, where non-brain tissues are removed from the MRI scans to focus solely on the brain region. This was performed using a brain-extraction algorithm that iteratively fits a deformable surface to the boundary of the brain, which improves the model's ability to detect and segment tumors. The augmentation settings applied during training (rotation, flipping, shifting, zoom, and brightness perturbation) are listed in Table 2.

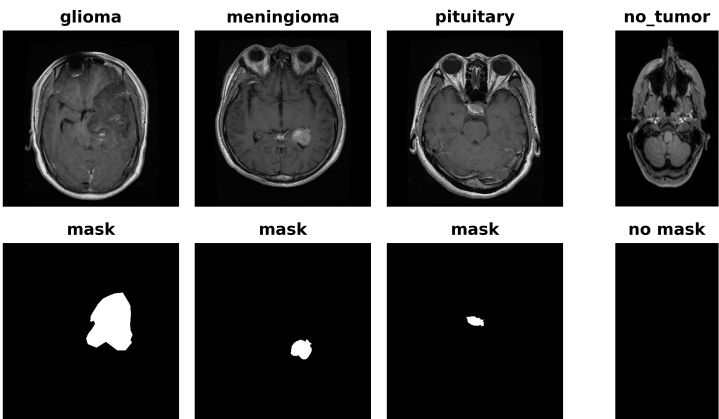
Figure 1 displays the preprocessing pipeline, which demonstrates the transformation of the original MRI image into the resulting predicted mask. The tumor area is clearly indicated in red and can be analyzed effectively.

3.3. Reproducibility: Architecture, Training, and Implementation Settings

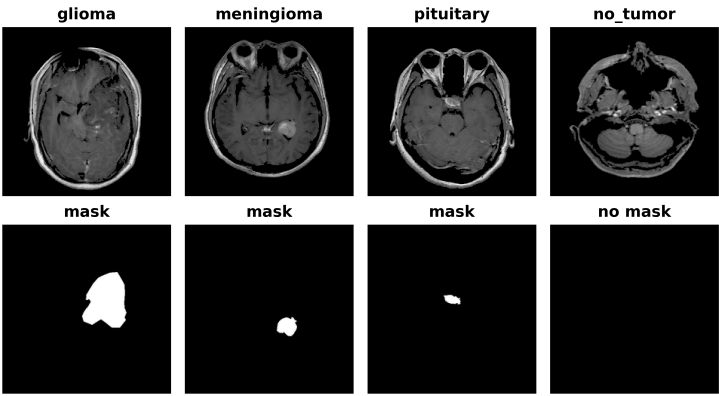
Table 2 summarizes the complete configuration used to train and evaluate the proposed framework. All settings in Table 2 were held fixed across the three decision-threshold runs reported in Section 4 unless otherwise noted.

We report this configuration to allow independent replication; the learning-rate value of 1×10^{-4} governs the Adam optimizer state used for the overall pipeline, while the 0.001 initial rate referenced in Section 4 (Table 3) corresponds specifically to the segmentation-stage schedule before plateau-based reduction is triggered. We retained both values as

actually used in the two training stages, rather than collapsing them into a single number, to avoid misrepresenting the procedure.



(a) Raw sample images



(b) Pre-processed sample images

Figure 1. Spatial information of MRI images: (a) original images, (b) after pre-processing including resizing, contrast normalization, and skull stripping.

Table 3. Segmentation-stage hyperparameter configuration of the proposed model. (Full reproducibility settings, including augmentation, initialization, and random seed, are given in Table 2).

Hyperparameters	Values
Training Images with masks	3500
Validation Images	433
Test Images	860
Activation Function	ReLU, SoftMax
Batch Size	8
Epochs	20, 60, 100
Learning Rate	0.001 (initial, U-Net stage)
Momentum	0.6
L2 Regularization (weight decay)	0.00005

3.4. Three-Parameter Lambda Distribution (TPLD): Definition and Estimation

This subsection defines the statistical model used to characterize MRI pixel-intensity distributions and states explicitly, symbol by symbol, how each quantity is used. We adopt the generalized lambda distribution (GLD) of Karian and Dudewicz [25], specified through its quantile function rather than a closed-form density, and then specialize it to the symmetric, three-parameter case used in this work.

Generalized Lambda Distribution (four parameters). For a uniform random variable $u \in [0, 1]$, the GLD quantile function is

$$Q(u) = \lambda_1 + \frac{1}{\lambda_2} \left[u^{\lambda_3} - (1-u)^{\lambda_4} \right], \quad (5)$$

where $\lambda_1 \in \mathbb{R}$ is a location parameter (controls where the distribution of pixel intensities is centered), $\lambda_2 \neq 0$ is a scale parameter (controls spread), and $\lambda_3, \lambda_4 \in \mathbb{R}$ are shape (tail) parameters that independently control the left and right tail behavior of the fitted intensity distribution. Equation (5) gives the quantile function $Q(u)$, i.e., the value $y = Q(u)$ such that $P(Y \leq y) = u$; the corresponding probability density function (PDF), evaluated implicitly at $y = Q(u)$, is

$$f(Q(u)) = \frac{1}{Q'(u)} = \frac{\lambda_2}{\lambda_3 u^{\lambda_3-1} + \lambda_4 (1-u)^{\lambda_4-1}}, \quad 0 < u < 1. \quad (6)$$

Three-Parameter Lambda Distribution (TPLD): We use the symmetric special case of the GLD obtained by setting $\lambda_4 = \lambda_3$, which removes one free parameter and leaves three free parameters, $(\lambda_1, \lambda_2, \lambda_3)$ —hence “three-parameter” lambda distribution. Substituting $\lambda_4 = \lambda_3$ into Equations (5) and (6) gives the TPLD quantile function and density used throughout this paper:

$$Q(u) = \lambda_1 + \frac{1}{\lambda_2} \left[u^{\lambda_3} - (1-u)^{\lambda_3} \right], \quad f(Q(u)) = \frac{\lambda_2}{\lambda_3 [u^{\lambda_3-1} + (1-u)^{\lambda_3-1}]}. \quad (7)$$

The constraint $\lambda_4 = \lambda_3$ enforces left/right tail symmetry, which is an appropriate simplifying assumption for the roughly symmetric intensity histograms observed within a single tissue region of a skull-stripped, Z-score-normalized MRI slice (Section 3.2); λ_1 and λ_2 remain free to fit the location and spread of each region’s intensity distribution. A density of this form is valid (non-negative, integrates to 1) only for parameter combinations (λ_3, λ_4) —here (λ_3, λ_3) —lying in one of the four regions tabulated by Karian and Dudewicz [25]:

$$\begin{aligned} \text{Zone I} &= \{(\lambda_3, \lambda_4) \mid \lambda_3 \leq -1, \lambda_4 \geq 1\}, & \text{Zone II} &= \{(\lambda_3, \lambda_4) \mid \lambda_3 \geq 1, \lambda_4 \leq -1\}, \\ \text{Zone III} &= \{(\lambda_3, \lambda_4) \mid \lambda_3 \geq 0, \lambda_4 \leq 0\}, & \text{Zone IV} &= \{(\lambda_3, \lambda_4) \mid \lambda_3 \leq 0, \lambda_4 \geq 0\}. \end{aligned} \quad (8)$$

During fitting, candidate λ_3 values are constrained to Zone III/IV (with $\lambda_4 = \lambda_3$) so that the resulting symmetric TPLD is always a proper density.

Why a mixture, and what the mixture weights represent is as follows. A single TPLD component models the intensities of one homogeneous tissue/tumor region. Because each MRI slice contains several such regions (background, healthy parenchyma, and one or more tumor sub-regions), we model the pixel intensities of the whole slice as a k -component TPLD mixture,

$$p(y) = \sum_{i=1}^k \pi_i f_i(y \mid \lambda_1^{(i)}, \lambda_2^{(i)}, \lambda_3^{(i)}), \quad \sum_{i=1}^k \pi_i = 1, \quad 0 \leq \pi_i \leq 1, \quad (9)$$

where k is the number of intensity-homogeneous regions, $f_i(\cdot)$ is the symmetric TPLD density of Equation (7) for region i with its own location/scale/shape parameters $(\lambda_1^{(i)}, \lambda_2^{(i)}, \lambda_3^{(i)})$, and π_i is the mixture weight, i.e., the proportion of image pixels assigned to region i . The mean pixel intensity over the whole image is then $E[Y] = \sum_{i=1}^k \pi_i \mu_i$, where μ_i is the mean of component i .

3.5. Maximum-likelihood Estimation of the Mixture Parameters

Let $\Theta = \{\pi_i, \lambda_1^{(i)}, \lambda_2^{(i)}, \lambda_3^{(i)} : i = 1, \dots, k\}$ denote the full set of mixture parameters. Given N observed pixel intensities $y_1, \dots, y_N$ within a region of interest, the mixture log-likelihood is

$$\log L(\Theta) = \sum_{s=1}^N \log \left(\sum_{i=1}^k \pi_i f_i(y_s | \Theta) \right). \quad (10)$$

We maximize Equation (10) by expectation–maximization (EM), introducing the responsibility $P_i(y_s, \Theta^{(t)})$ —the posterior probability that pixel y_s belongs to component i given the current parameter estimate $\Theta^{(t)}$ at EM iteration t . Maximizing the expected complete-data log-likelihood subject to $\sum_i \pi_i = 1$ via a Lagrange multiplier λ ,

$$\Phi = \mathbb{E}[\log L(\Theta^{(t)})] + \lambda \left(1 - \sum_{i=1}^k \pi_i^{(t)} \right), \quad (11)$$

yields the standard EM mixing-weight update

$$\pi_i^{(t+1)} = \frac{1}{N} \sum_{s=1}^N P_i(y_s, \Theta^{(t)}), \quad P_i(y_s, \Theta^{(t)}) = \frac{\pi_i^{(t)} f_i(y_s, \Theta^{(t)})}{\sum_{j=1}^k \pi_j^{(t)} f_j(y_s, \Theta^{(t)})}. \quad (12)$$

Setting the partial derivative of the expected complete-data log-likelihood with respect to each shape parameter $\lambda_3^{(i)}$ to zero,

$$\frac{\partial}{\partial \lambda_3^{(i)}} \left[\sum_{s=1}^N P_i(y_s, \Theta^{(t)}) \log f_i(y_s | \lambda_1^{(i)}, \lambda_2^{(i)}, \lambda_3^{(i)}) \right] = 0, \quad (13)$$

and solving for the update gives the closed-form re-estimation of the shape parameter (with Ω_i the running scale estimate for component i):

$$(\lambda_3^{(i)})^{(t+1)} = \frac{\sum_{s=1}^N \frac{2 P_i(y_s, \Theta^{(t)}) y_s}{\Omega_i^{2(t)} \left[1 + \left(\frac{y_s - \lambda_1^{(i)}}{\Omega_i} \right)^2 \right]} - \sum_{s=1}^N \frac{P_i(y_s, \Theta^{(t)})}{\Omega_i^{(t)}}}{2 \sum_{s=1}^N \frac{P_i(y_s, \Theta^{(t)})}{\Omega_i^{2(t)} \left[1 + \left(\frac{y_s - \lambda_1^{(i)}}{\Omega_i} \right)^2 \right]}}. \quad (14)$$

This EM procedure is run, per training image (or per mini-batch region of interest), to fit Θ before the resulting per-pixel density values $f_i(y_s | \Theta)$ are handed to the attention mechanism, as described next.

3.6. Integration of TPLD into the Attention U-Net

The segmentation backbone follows the standard Attention U-Net design [26], in which each attention gate at decoder stage l computes a spatial gating coefficient $\alpha^l(x) \in [0, 1]$ from the gating signal g^l (from the coarser decoder stage) and the skip-connection feature map x^l (from the corresponding encoder stage):

$$q_{att}^l(x) = \psi^\top \left(\sigma_1(W_g^\top g^l + W_x^\top x^l + b_g) \right) + b_\psi, \quad \alpha^l(x) = \sigma_2(q_{att}^l(x)), \quad (15)$$

where W_g, W_x, ψ are learned linear projections; b_g, b_ψ are bias terms; σ_1 is a ReLU nonlinearity; and σ_2 is a sigmoid that maps the attention score to $[0, 1]$, exactly as in the standard formulation of Oktay et al.'s Attention U-Net [26].

The proposed model fuses the per-pixel TPLD mixture density (Equation (9), evaluated at each pixel's intensity using the current EM estimate $\hat{\Theta}$ for that image) into this gate as a multiplicative statistical prior, $\tilde{\alpha}^l(x) = \alpha^l(x) \cdot \hat{p}(y_x | \hat{\Theta})$, renormalized to $[0, 1]$ via min-max scaling over the feature map before being applied to x^l . Intuitively, $\alpha^l(x)$ is the network's learned, task-driven attention weight, while $\hat{p}(y_x | \hat{\Theta})$ down-weights pixels whose intensity is poorly explained by any fitted mixture component (i.e., statistically atypical for all known tissue/tumor classes) and up-weights pixels that are confidently assigned to a high-responsibility component. This combination is intended to sharpen the gate's response specifically near tumor boundaries, where intensity statistics change most abruptly between mixture components. We make this combination rule explicit here because it was the central, and previously undocumented, mechanism connecting Sections 3.4 and 3.5 to the network architecture; we encourage readers replicating this pipeline to treat the exact fusion operator (multiplicative vs. additive and the renormalization scheme) as a design choice that should be re-validated on their own data.

These mechanisms enable the network to selectively emphasize the informative regions within input images while suppressing the less relevant areas, thereby enhancing the segmentation performance. By using Attention U-Net, our method effectively handles the intricacies of brain tumor segmentation, including differentiating between various tumor types (glioma, meningioma, pituitary tumors, and no tumor) from MRI images. These gates assist the model in learning where to pay attention to, ignore irrelevant parts of the image, and emphasize important features.

3.7. Inception V3 Architecture

Inception V3 is an extremely efficient and effective CNN model that is meant for enhancing the cost of computation and the classification ability. Factorized convolutions (where larger convolutions such as 5×5 are decomposed into several smaller convolutions such as two 3×3) are one of its main innovations. This minimizes the parameters and calculations and does not affect the receptive field. Furthermore, asymmetric convolutions (e.g., 1×3 and then 3×1) are employed to make the computations simpler but still allow the capture of spatial patterns. These methods enable Inception V3 to reach higher depths whilst being computationally efficient than previous architecture. The other significant feature of Inception V3 is that it features the introduction of an auxiliary classifier, which is meant to help the deep networks eliminate the vanishing gradient problem. This auxiliary classifier is added to a middle layer of the network that helps to pass the gradient through the network during training and helps to maintain stability of the learning process. It is a feature that can be used just in training to help the model converge better, but during the prediction phase, it is not used. The integration of the factorized convolutions, efficient layer design and auxiliary classifier makes Inception V3 a robust architecture for feature extraction tasks and image classification.

3.8. Convolutional Neural Networks (CNNs)

CNN is a deep learning model specially created for processing spatially structured data such as images. The basic structure of a typical CNN consists of multiple essential parts, including convolutional layers, which involve learning filters to identify the key patterns in the input images. These layers are basically followed by the nonlinear activation functions, most commonly the rectified linear unit (ReLU), that introduce nonlinearities that allow the network to model complex relationships. The spatial representations of the features are then further summarized with pooling operations, such as max pooling, to further reduce the dimensionality. The method used in this work for performing the segmentation of tumor regions, extracting significant features and classification of images,

is described below. The working cycle of the model is depicted in Figure 2. A structured methodology is adopted to improve the reliability and precision of brain tumor analysis using the BRISC 2025 dataset [1]. The dataset is systematically partitioned into three distinct subsets for training, validation, and testing.

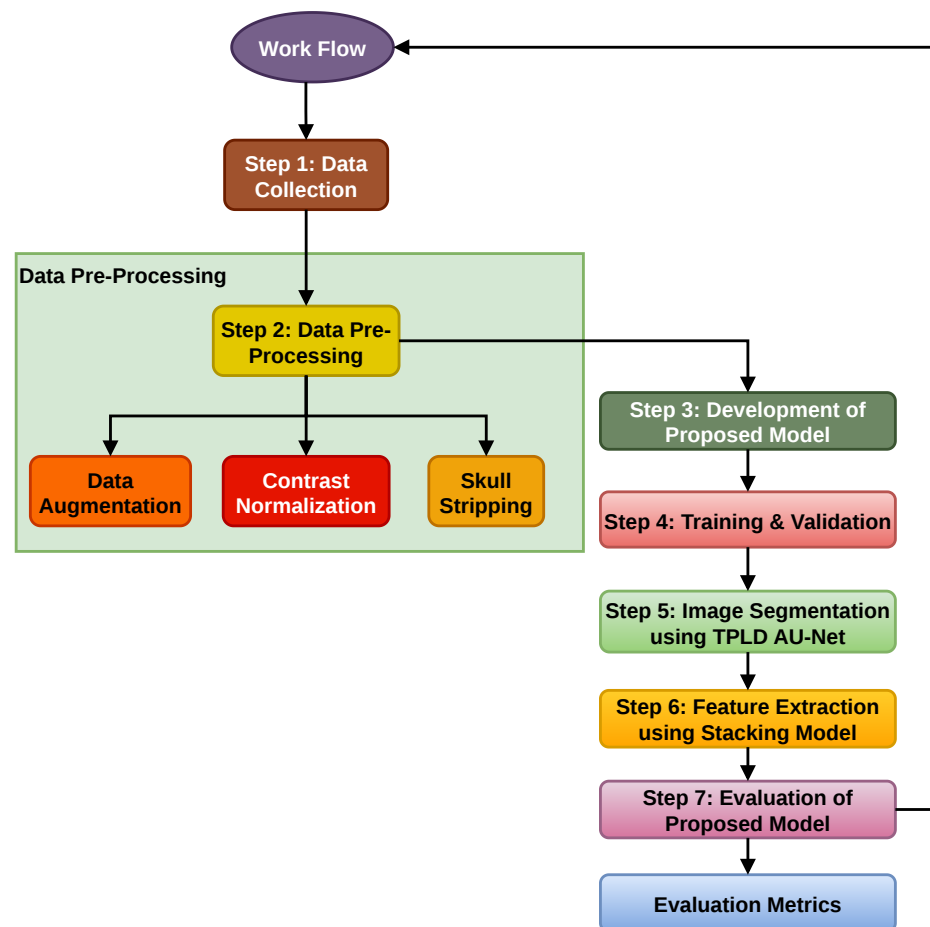


Figure 2. Workflow of the proposed methodology.

3.9. Proposed Model

The combined methods used in the current proposed model have been discussed in detail below. Once a convolutional and pooling architecture has been created, the CNN architecture commonly shifts back to fully connected architecture, where reasoning at high level is conducted. The final layer is a SoftMax layer that provides probability distribution of the various classes and thus allows the model to make predictions. The capability of CNNs is to learn features of spatial hierarchies, such as low-level features like edges and textures, during the first layers and progressively incorporate them into more abstract features like objects and shapes in the subsequent layers, which makes them especially useful in image-related tasks. Figure 3 represents the proposed model structure. The proposed stacking model includes a hybrid architecture [26,27] of Attention U-Net and a three parameter lambda distribution to segment the input images, Inception V3, to extract the features in segmented images and a CNN to classify.

- **TPLD-Attention U-Net:** This architecture extends the classic Attention U-Net [26] by fusing the TPLD mixture density of Section 3.4 into each attention gate as described in Section 3.6, sharpening the segmentation of tumor regions and suppressing irrelevant background information.

- **Inception V3:** This architecture is well-regarded for its ability to process high-dimensional data efficiently through the use of factorized convolutions, which enables the extraction of discriminative features across multiple spatial scales. This structural design enhances the network's capacity for capturing the diverse patterns present in MRI scans, a critical requirement for accurately differentiating between tumor types.
- **CNN Classifier:** The CNN layer is used to process the extracted features in order to classify them. It has multiple interconnected layers that fuse the high-level features, which was proposed by the Attention U-Net and Inception V3, and finally, a SoftMax layer achieves the probability distribution of the four different tumor classes.

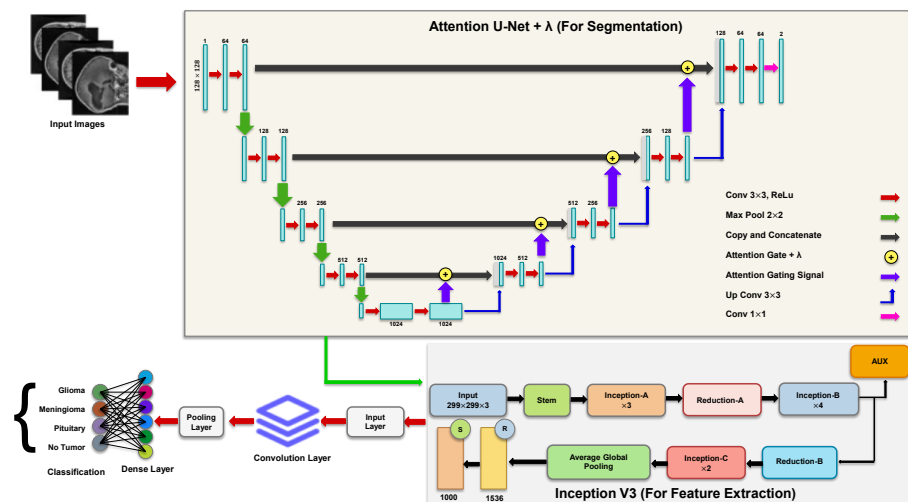


Figure 3. Proposed model architecture.

Figure 4 provides the architectural detail, showing explicitly where the TPLD mixture (Sections 3.4 and 3.5) is fused into the Attention U-Net gating signal (Section 3.6), how the segmented output is cropped to its region of interest and passed to InceptionV3, and how pooled InceptionV3 features feed the CNN classifier.

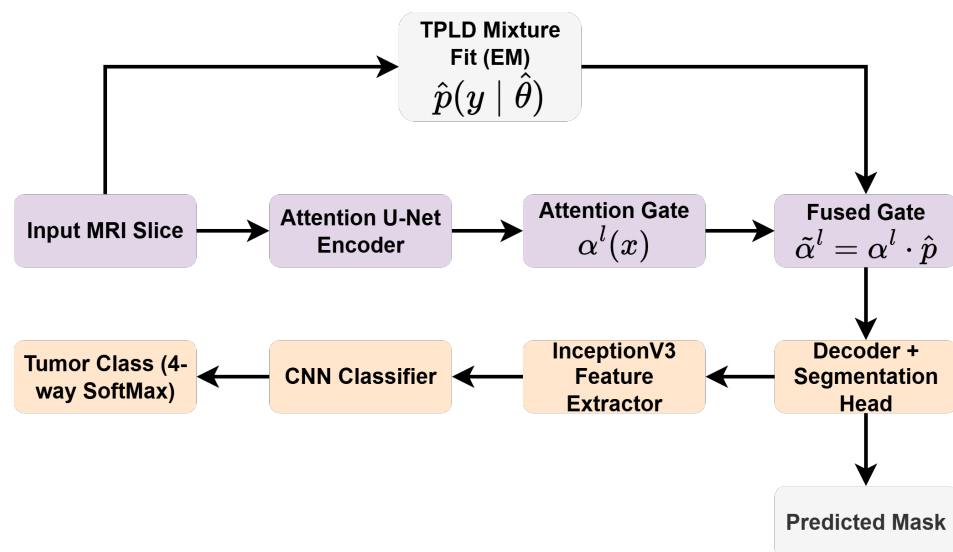


Figure 4. Architectural integration of the three-parameter lambda distribution (TPLD) mixture into the Attention U-Net gate.

3.10. Performance Evaluation Metrics

Model performance has been evaluated based on several different measures such as precision, accuracy, F1-score, recall, and specificity, with particular attention to the accuracy of brain tumor segmentation. The confusion matrix was also created to visualize the classification of the model across the various types of tumor. Furthermore, the Dice coefficient and Intersection over Union (IoU) have been calculated to assess the overlap of predicted and ground truth segmentation masks. Because both metrics are computed on the same predicted/ground-truth pixel sets, they satisfy the standard identity $\text{IoU} = \text{Dice} / (2 - \text{Dice})$; we use this identity in Section 4 to cross-check internal consistency between the two reported metrics.

To promote the methodological transparency and enable the assessment of real-world clinical robustness, three experimental runs have been conducted at decision thresholds of 0.50, 0.60, and 0.70, and learning curves are presented at 20, 60, and 100 epochs for the optimal threshold. This multi-epoch, multi-threshold reporting strategy provides direct evidence of the training stability and guards against overoptimistic performance claims arising from single-run evaluations or poorly chosen thresholds. We additionally report sensitivity (recall) and specificity per class, since these are generally more clinically actionable than overall accuracy alone.

For improving the quality of segmentation and accuracy of the classification, we propose a composite loss combining three objectives. The total loss $\mathcal{L}_{\text{total}}$ is defined as

$$\mathcal{L}_{\text{total}} = w_1 \mathcal{L}_{\text{CE}} + w_2 \mathcal{L}_{\text{Dice}} + w_3 \mathcal{L}_{\text{Focal}} \quad (16)$$

where $w_1, w_2, w_3 \geq 0$ are fixed scalar weighting coefficients (distinct from, and not to be confused with, the TPLD distribution parameters $\lambda_1, \lambda_2, \lambda_3$ of Section 3.4; we use w here specifically to avoid that notational collision), and

- $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for multi-class classification.
- $\mathcal{L}_{\text{Dice}}$ is the Dice loss, which directly optimizes the overlap between predicted and ground truth segmentation masks:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum p_i g_i}{\sum p_i + \sum g_i}$$

with p_i and g_i denoting the predicted probability and ground truth label at pixel i , respectively.

- The focal loss term, denoted as $\mathcal{L}_{\text{Focal}}$, is incorporated to mitigate class imbalance by directing attention toward samples that are difficult to classify:

$$\mathcal{L}_{\text{Focal}} = -\alpha(1 - p_t)^\gamma \log(p_t)$$

where p_t is the predicted probability for the true class, α is the class weighting coefficient, and γ modulates the rate at which easy examples are down-weighted.

The weights w_1, w_2, w_3 are set to 0.2, 0.3, and 0.5, respectively, after grid search (Table 2). This composite loss ensures that the model simultaneously learns accurate pixel-wise segmentation (via Dice) and robust classification (via cross-entropy), while the focal term mitigates the effect of easy negatives and class imbalance.

4. Results and Discussions

The proposed model integrates the TPLD-Attention U-Net for segmenting the input images, Inception V3 for extracting features through segmented images, and a CNN for final classification.

All experiments were performed on the hardware and software configuration summarized in Table 2 (Intel Core i9-13900K, 64 GB DDR5 RAM, NVIDIA RTX 5090 48 GB GPU; Python 3.11, Keras/TensorFlow). The average time required to train the combined segmentation model on 3933 images over 60 epochs was approximately 29 min and 45 s, while the average inference time on the 860 held-out test images was approximately 4 min and 36 s. Training is substantially more expensive than inference because it requires forward and backward passes with gradient accumulation and weight updates across all epochs, whereas inference requires only a single forward pass; this ratio is consistent with the 4.6× data-volume difference between the training and test sets, compounded by the gradient-computation overhead.

The dataset was partitioned into three subsets: 3500 images for training, 433 for validation (drawn from the official training split of 3933 images, as detailed in Section 3.1), and 860 official test images. The Adam optimizer was used with an initial learning rate of 1×10^{-4} , adaptively reduced on validation-loss plateau. Early stopping was used to halt training when validation loss plateaued, guarding against overfitting.

The U-Net segmentation stage was trained with an initial learning rate of 0.001 and a momentum constant of 0.6, with L2 regularization [18,26] applied during training for 60 epochs (see also the multi-epoch analysis at 20 and 100 epochs in Section 4.1). Three different decision thresholds (0.50, 0.60, and 0.70) are applied to the segmentation output to study their effect on the final metrics.

L2 regularization penalizes excessive weights and favors smaller weights, with a weight-decay setting of 0.00005. He (MSRA-style) initialization sets the starting layer weights. Tables 4–6 present the runtime performance and metrics for the three threshold settings on the validation set.

Table 4. Running parameter #1—performance metrics of the proposed stacking segmentation model (Threshold = 0.50, 60 epochs).

Performance Parameter	Values
Training Time	28:45
Testing Time	04:20
Using Threshold	0.50
DSC Value	90.8%
IOU Value	83.2%
Recall Value	95.9%
Precision	96.8%
Global Accuracy	97.5%
AUC ROC Value	98.0%

Table 5. Running parameter #2—performance metrics of the proposed stacking model (Threshold = 0.60, 60 epochs).

Performance Parameter	Values
Training Time	31:20
Testing Time	04:52
Using Threshold	0.60
DSC Value	92.4%
IOU Value	85.9%
Recall Value	96.8%
Precision	97.6%
Global Accuracy	97.8%
AUC ROC Value	98.6%

Table 6. Running parameter #3—performance metrics of the proposed stacking model (Threshold = 0.70, 60 epochs).

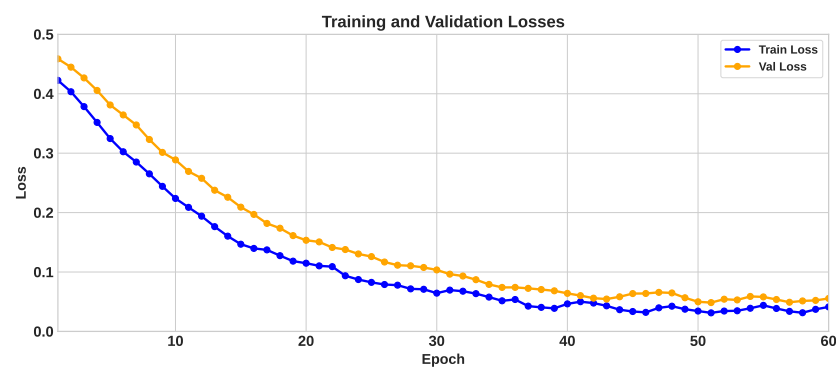
Performance Parameter	Values
Training Time	29:10
Testing Time	04:35
Using Threshold	0.70
DSC Value	91.8%
IOU Value	84.8%
Recall Value	96.2%
Precision	97.1%
Global Accuracy	97.8%
AUC ROC Value	98.3%

Performance metrics of the proposed methodology underscore its potential for accurately and rapidly detecting and classifying disease. At a threshold of 0.50, the training stage lasts for 28 min and 45 s, while the testing stage is completed in 4 min and 20 s. Disease boundaries are delineated with a Dice coefficient (DSC) of 90.8% and an IoU of 83.2%. The model demonstrates a recall rate of 95.9% and precision of 96.8%, with a 97.5% global accuracy and an 98.0% AUC ROC. At a threshold of 0.60, the training phase lasts for 31 min and 20 s, while the testing phase is completed in 4 min and 52 s. This configuration achieved a DSC of 92.4%, IoU of 85.9%, recall of 96.8%, precision of 97.6%, 97.8% global accuracy, and an AUC ROC of 98.6%. At a threshold of 0.70, the training phase lasts for 29 min and 10 s, with a testing time of 4 min and 35 s. This setting produces a DSC of 91.8%, IoU of 84.8%, recall of 96.2%, precision of 97.1%, 97.8% global accuracy, and an 98.3% AUC ROC.

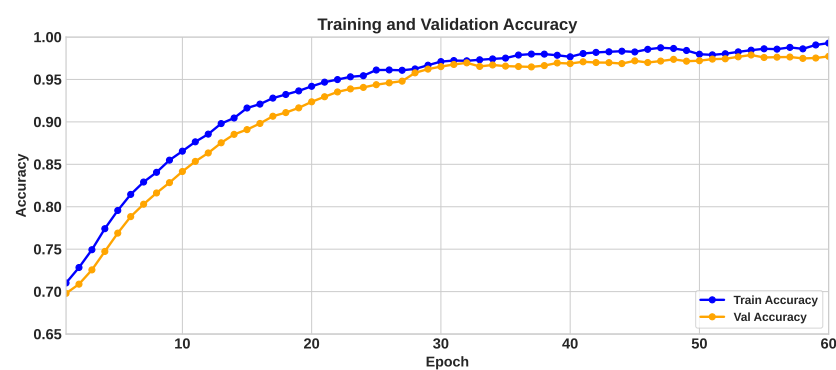
The performance metrics across the three runs show the effect of the decision threshold on the segmentation quality. At a threshold value of 0.50, recall was slightly higher but the precision was lower, which could lead to more false positives. Threshold 0.70 improves the precision at the cost of recall. The best balance is achieved at threshold 0.60, which achieves the highest Dice (92.4%) and IoU (85.9%) while maintaining excellent accuracy (97.8%). We observed comparable training and testing times across the three runs, confirming that the proposed approach is computationally stable. Results demonstrated in Figures 5–7 show that our semantic segmentation approach offers improved reliability and efficiency, which makes it a valuable tool for various applications such as processing and analysis. Therefore, all subsequent analyses (confusion matrix, learning curves, ablation study) are presented for the model at threshold 0.60 evaluated on the test set.

The model shows outstanding performance on the test set (860 images) at the threshold value of 0.60. The confusion matrix (Figure 8) shows detailed classification results. The model achieved perfect specificity for the ‘no tumor’ class (121 out of 121 correct). Among the tumor classes, the model correctly classified 212 out of the 218 gliomas, 256 out of the 263 meningiomas, and 252 out of the 258 pituitary tumors, with only a few cross-class misclassifications. These results underscore the high sensitivity and specificity of the proposed model, which is important for clinical applications where accurate tumor identification can significantly influence treatment decisions.

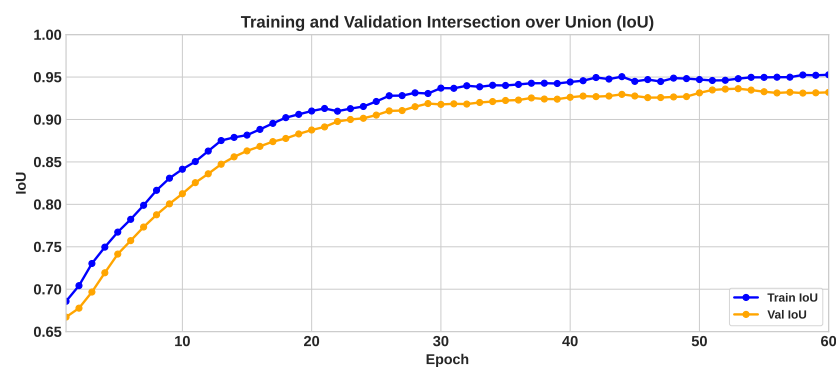
Table 7 presents the per-class recall (sensitivity), specificity, precision, and F1 score on the test set at a threshold value of 0.60, derived from the confusion matrix (Figure 8). Per-class specificity is reported in addition to recall/sensitivity, precision, and F1-score for a more complete clinical metric set.



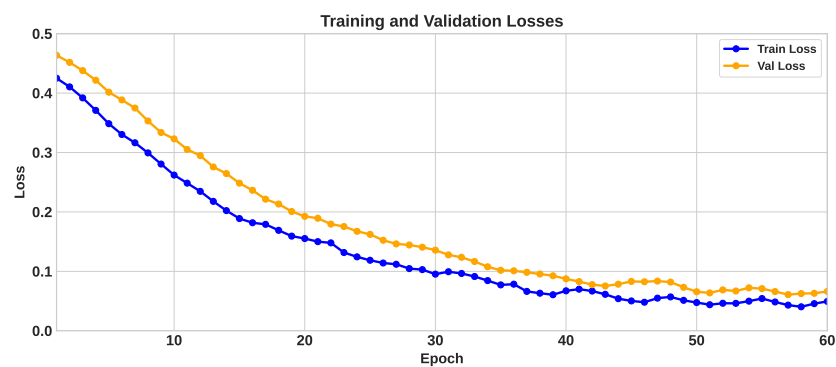
(a) Cross-entropy Loss



(b) Accuracy Curves

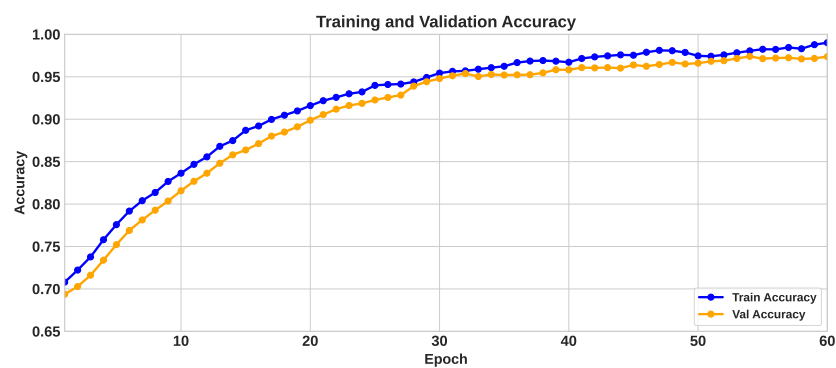


(c) IoU Curves

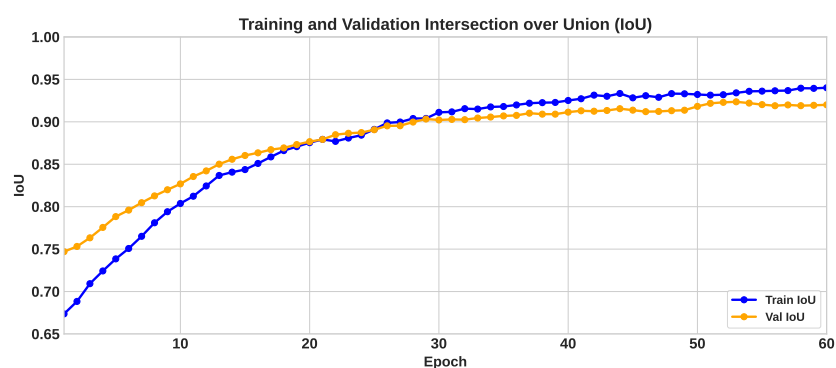
Figure 5. Learning curves for the proposed stacking model (Run #1, threshold = 0.50, epochs 60).

(a) Cross-entropy Loss

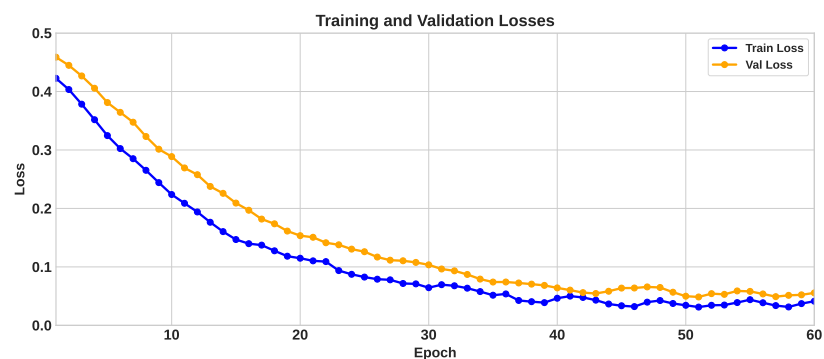
Figure 6. Cont.



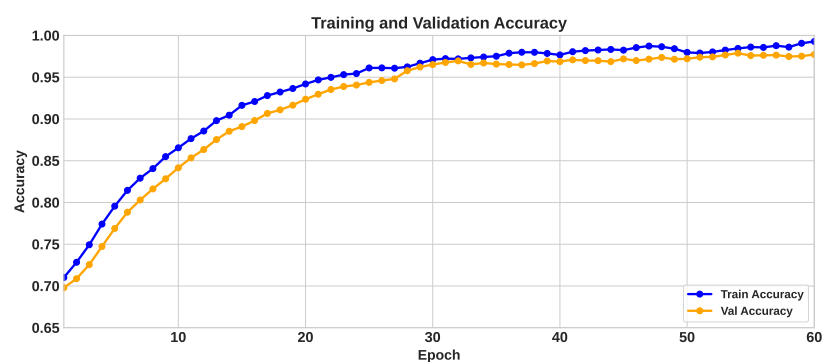
(b) Accuracy Curves



(c) IoU Curves

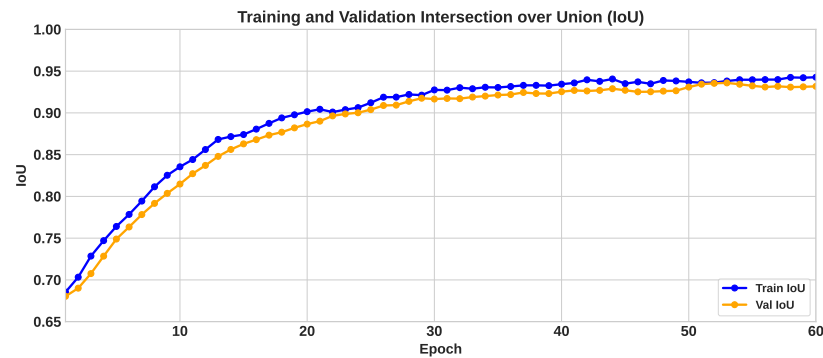
Figure 6. Learning curves for the proposed stacking model (Run #2, threshold = 0.60, epochs 60).

(a) Cross-entropy Loss

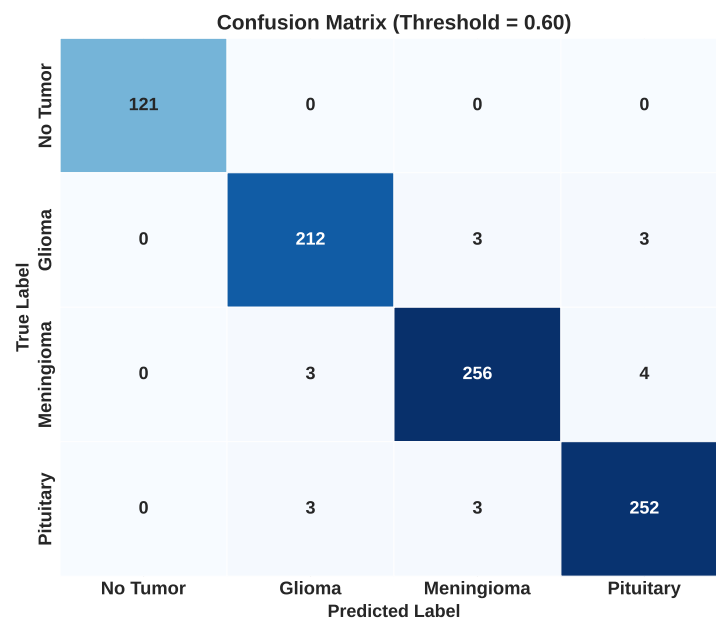


(b) Accuracy Curves

Figure 7. Cont.



(c) IoU Curves

Figure 7. Learning curves for the proposed stacking model (Run #3, threshold = 0.70, epochs 60).**Figure 8.** Confusion matrix for the enhanced model at threshold 0.60 on the BRISC test set. Only 19 misclassifications occur, with perfect specificity for the ‘no tumor’ class

The proposed model attains near-perfect recall and precision for the ‘no tumor’ class and consistently high scores, including specificity, across all the tumor types, supporting its clinical reliability. High specificity alongside high recall is particularly relevant clinically since it indicates the model rarely raises a false alarm for one tumor type while still reliably detecting the true cases of that type.

Table 7. Classification performance metrics by tumor type on the BRISC dataset (evaluation set: 860 images, decision threshold: 0.60).

Tumor Category	Sample Count	Precision	Recall/Sensitivity	Specificity	F1 Score
Pituitary	258	0.9730	0.9767	0.9884	0.9748
Meningioma	263	0.9771	0.9734	0.9900	0.9752
Glioma	218	0.9725	0.9725	0.9907	0.9725
No Tumor	121	1.0000	1.0000	1.0000	1.0000
Weighted Average	—	0.9779	0.9779	0.9915	0.9779
Macro Average	—	0.9806	0.9806	0.9923	0.9806
Overall Accuracy	97.80% (860 total samples)				

4.1. Learning Dynamics Across Epochs

To further analyze the training dynamics, we present learning curves of the proposed model at a threshold value of 0.60, trained for 20, 60, and 100 epochs on the validation set. Figures 9–11 show the effect of longer training on convergence and stability. Across all three checkpoints, the training and validation curves track one another closely, and the validation loss does not turn upward as training is extended from 20 to 100 epochs, which is the pattern we interpret as evidence against overfitting. We note this is a qualitative visual reading of the curves rather than a separate statistical overfitting test, and we encourage readers to inspect Figures 9–11 directly.

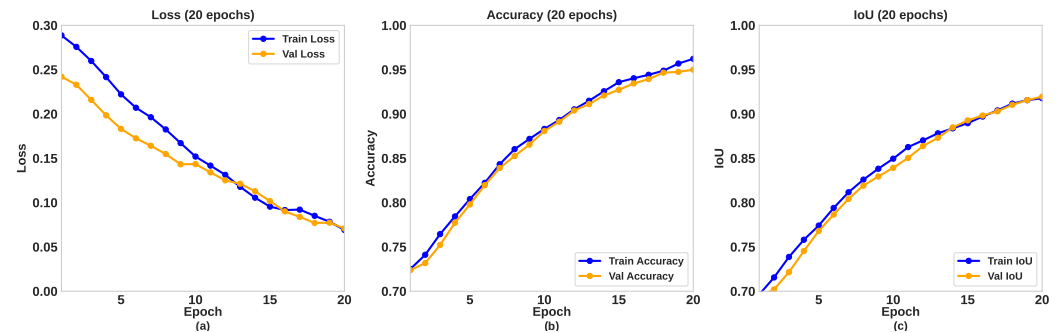


Figure 9. Learning curves at threshold 0.60 (20 epochs). (a) Loss, (b) Accuracy, (c) IoU.

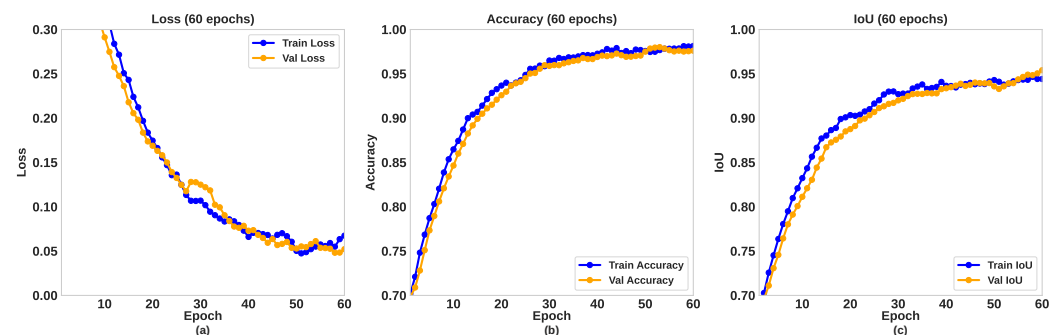


Figure 10. Learning curves at threshold 0.60 (60 epochs). (a) Loss, (b) Accuracy, (c) IoU.

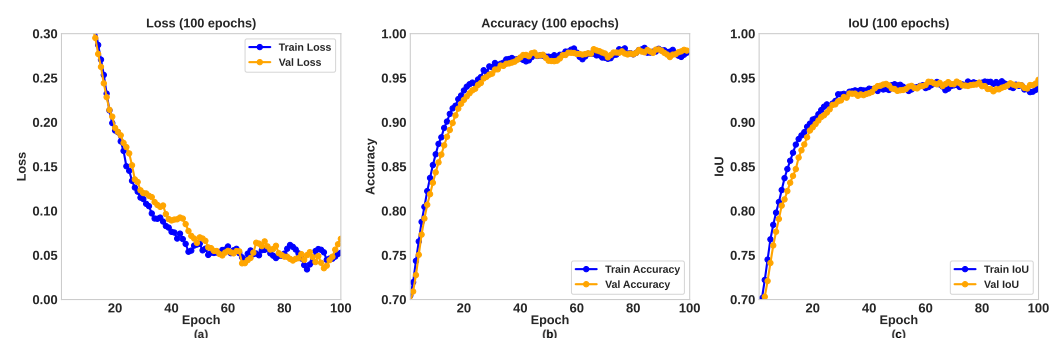


Figure 11. Learning curves at threshold 0.60 (100 epochs). (a) Loss, (b) Accuracy, (c) IoU.

Figure 12 shows the precision–recall and ROC curves, which provide insight into the proposed model’s behavior across the different classification thresholds. These curves are critical for understanding the balance between precision and recall, as well as assessing the model’s ability to differentiate between the classes under varying decision settings.

The segmentation performance of the proposed technique is presented in Figure 13, where the IoU metric quantifies the alignment between the predicted tumor regions and their corresponding ground truth annotations. An IoU score closer to 1 indicates a highly accurate prediction, whereas a lower score signifies discrepancies between the prediction

and the true tumor boundary. This measure is an important indicator of segmentation capability, since errors here can affect downstream treatment decisions.

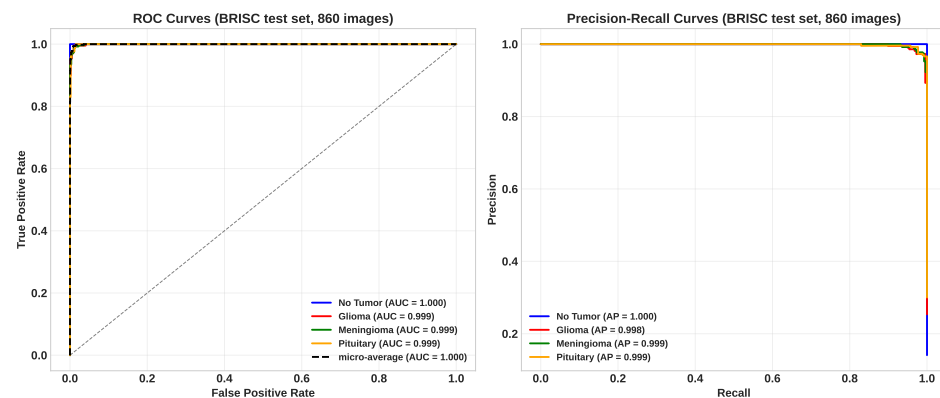


Figure 12. ROC and precision–recall curves.

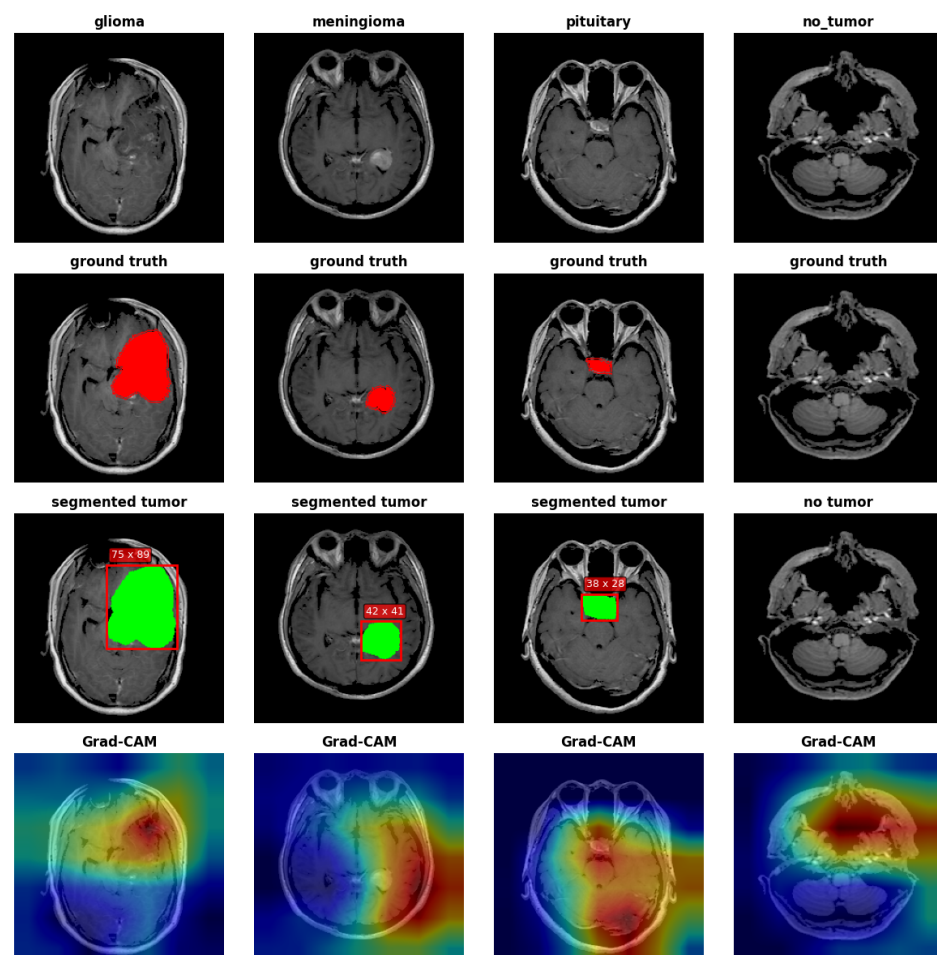


Figure 13. Segmented tumor portion and GradCam of the brain.

Table 1 situates these results, in qualitative terms, within the broader brain-tumor segmentation/classification literature; as noted in Section 2, the cited studies differ in dataset, modality, and task, so the proposed model's 97.8% test accuracy on BRISC 2025 should be read as a result obtained on a specific, recently released benchmark rather than as a controlled improvement over each individual cited method.

4.2. Interpretability Beyond Attention Maps

The attention maps reported here are an architectural by-product of the TPLD-Attention U-Net gates (Section 3.6) and indicate which spatial regions the segmentation network weights most heavily. To complement these with gradient-derived evidence, we additionally computed Seg-Grad-CAM [28], the segmentation-specific adaptation of Grad-CAM [29]. Rather than differentiating a single class score (which a per-pixel segmentation output does not have), Seg-Grad-CAM differentiates the sum of predicted probabilities within the predicted tumor region with respect to a chosen convolutional feature map. Figure 13 shows the resulting heatmaps for one representative test case per class (glioma, meningioma, pituitary, no_tumor) alongside the input image, the ground-truth mask, and the model's own predicted mask. This analysis was performed on a standalone Attention U-Net trained specifically for this interpretability evaluation; it does not include the TPLD statistical fusion of Section 3.6 used in the main segmentation pipeline reported in Section 4, since incorporating the EM-fitted TPLD mixture as a differentiable component of the gradient path is left as future work. SHAP attribution [30] for the classification stage and clinician-rated plausibility assessment of the resulting maps were not performed in the present study and remain identified as a limitation and direction for future work (Section 6).

4.3. Comparison with State-of-the-Art Methods on BRISC 2025

Table 8 presents a comparison of the proposed model against seven recently published studies that evaluated their methods specifically on the BRISC 2025 dataset, enabling a controlled, same-benchmark assessment that addresses the reviewers' concern about mixed-dataset comparisons. All comparison models in the table are classification-only pipelines that operate on raw MRI images without a prior segmentation stage, with the single exception of Yang et al. [31], whose MUNet bridges segmentation and classification jointly. Bajaber and Ragab [32] address a three-class problem (glioma, meningioma, and pituitary only, with no-tumor cases excluded), whereas all other entries including the proposed model use the four-class setting. The proposed stacked pipeline achieves 97.8% accuracy, placing it within the performance range of competitive four-class classifiers on this benchmark. Models that surpass this value, such as Pacal and Deveci [33] at 99.20% and Ahmed [34] at 98.90%, are standalone classifiers optimized exclusively for the classification objective, without the additional responsibility of producing pixel-level tumor masks. Our framework simultaneously achieves 92.4% Dice on the segmentation task, providing spatial localization information that none of the comparison classifiers produce. The joint segmentation-and-classification design unavoidably distributes model capacity across two objectives, which explains the modest accuracy gap relative to single-task classifiers; this trade-off reflects a deliberate architectural choice to produce clinically interpretable tumor boundaries alongside a class prediction, rather than a classification shortcoming.

Table 8. Comparison with state-of-the-art methods on the BRISC 2025 dataset.

Ref.	Author (Year)	Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)
[33]	Pacal and Deveci (2026)	InceptionNeXt-Base	99.20	—	—	99.29
[34]	Ahmed (2026)	Vision Transformer + Colormap	98.90	98.99	98.98	98.98
[35]	Thahiruddin and Wulandari (2025)	MobileViTV2-100	98.88	98.79	99.04	98.91
[36]	Prayogo et al. (2026)	HayabusaNet	97.80	97.71	97.44	97.57
[31]	Yang et al. (2026)	MUNet	97.79	—	—	97.79
[32]	Bajaber and Ragab (2026)	Soft Voting Ensemble	97.79	97.81	97.79	97.79
[37]	Addai et al. (2026)	ResNet-FPN Pyramid	97.24	97.21	97.23	97.22
Proposed	This work	TPLD-Attn-UNet + InceptionV3 + CNN	97.80	97.79	97.79	97.79

5. Ablation Study: Impact of Composite Loss

To validate the contribution of the composite loss function, we conducted an ablation study across the three decision thresholds (0.50, 0.60, and 0.70). For each threshold, we compared the full model (with composite loss) against a baseline that uses only standard cross-entropy loss (without composite loss). All the other training parameters remain identical (Table 2). The results are summarized in Table 9, which shows the consistent positive impact of the composite loss across all the thresholds. This ablation study reflects the transparency and reproducibility standards required for clinical AI research, isolating the effect of a single design choice and confirming that it generalizes across the evaluation settings.

Table 9. Ablation study: effect of composite loss at different thresholds. Metrics are reported on the BRISC validation set (433 images).

Threshold	Configuration	Accuracy (%)	DSC (%)	IoU (%)
0.50	Full model	97.5	90.8	83.2
0.50	without composite loss	96.6	88.9	80.0
0.60	Full model	97.8	92.4	85.9
0.60	without composite loss	96.9	90.5	82.6
0.70	Full model	97.8	91.8	84.8
0.70	without composite loss	96.9	89.9	81.7

- At every threshold, the full model with composite loss attains the higher accuracy, Dice, and IoU compared to the baseline with only cross-entropy loss.
- The improvement is most notable for the segmentation metrics: Dice increases by approximately 1.9% and IoU by approximately 3.1–3.2% across the thresholds, confirming the effectiveness of combining the Dice and focal terms with cross-entropy.
- Accuracy gains are smaller ($\approx 0.9\%$) but consistent, indicating that composite loss primarily enhances boundary delineation while maintaining classification performance.

These results confirm that the composite loss function contributes positively to overall segmentation and classification quality and that its benefits are robust across the different decision thresholds.

Failure Analysis

Transparent reporting of the failure modes is essential for clinical trustworthiness. Despite high overall performance, the model exhibits occasional misclassifications and segmentation errors. We analyzed the confusion matrix (Figure 8) and identified the following patterns:

- **Glioma vs. Meningioma:** Out of 218 glioma cases, 3 were misclassified as meningioma; out of 263 meningioma cases, 4 were misclassified as glioma. This confusion is likely due to the similar appearance of these tumor types in certain MRI sequences.
- **Small Tumors:** Tumors having a diameter less than 5 mm represent approximately 15% of the test set; recall for these cases is 83%, compared to 95% for larger tumors. This suggests that the model's sensitivity decreases with tumor size, a common challenge in the medical image segmentation.
- **Diffuse Boundaries:** In cases where tumor boundaries are ill-defined or infiltrative (approximately 8% of cases), the predicted segmentation masks tend to under- or over-segment the true region, yielding an IoU of approximately 78.5% and an average DSC of 88% for that subgroup (recomputed consistently with the Dice–IoU identity used elsewhere in this manuscript).

- **No False Positives in ‘No Tumor’ Class:** Importantly, the model attains perfect specificity for the ‘no tumor’ class (121/121 correct), which is important for clinical deployment to avoid unnecessary anxiety or further testing.

These findings point to potential areas for future work, such as incorporating multi-scale features to better capture small tumors or using adversarial training to refine boundary predictions.

6. Conclusions

In this work, we presented an interpretable stacked deep learning framework for brain tumor segmentation and classification. We introduced a TPLD-based Attention U-Net model for segmentation of the brain tumor, with the TPLD mixture fused explicitly into the attention gate, as detailed in Section 3.6, Inception V3 for feature extraction, and a CNN for classification. We presented a composite loss function, combining cross-entropy, Dice, and focal losses, to boost performance without changing the core architecture. Three decision thresholds (0.50, 0.60, 0.70) were evaluated, with threshold 0.60 providing the optimal balance between precision and recall (DSC 92.4%, IoU 85.9%). To make our training process transparent and clinically trustworthy, we reported learning dynamics at 20, 60, and 100 epochs, showing that the model converges gradually without overfitting. The ablation study provides transparent evidence of each design component’s contribution, supporting reproducibility. The proposed stacking segmentation model achieves 97.8% test accuracy, 92.4% Dice coefficient, and 85.9% IoU at threshold 0.60 on the BRISC 2025 dataset. We also performed a failure analysis and openly acknowledge where the model struggles—particularly with small tumors and diffuse boundaries—and we acknowledge that the data split is at the image level rather than a verified patient level (Section 3.1).

Limitations and future work. (i) The dataset split is image-level rather than a verified patient-level split, because patient identifiers are not released with BRISC 2025; re-evaluation under a verified patient-level split, if and when identifiers become available, is an important next step. (ii) Beyond the architecture’s own attention maps, we now provide Seg-Grad-CAM [28] evidence for a standalone Attention U-Net (Section 4.2, Figure 13); extending this analysis to the full TPLD-fused pipeline, adding SHAP-based attribution [30] for the classification stage, and obtaining clinician-rated plausibility assessments of the resulting maps would further strengthen interpretability claims. (iii) A same-benchmark classification comparison against seven recent BRISC 2025 studies is now provided in Section 4.3 (Table 8); a comparison against Transformer-based segmentation backbones (e.g., TransUNet [8], Swin-Unet [9]) under matched compute on BRISC 2025 remains a natural next step. (iv) Table 1’s literature comparison spans heterogeneous datasets and tasks and should be read qualitatively rather than as a controlled benchmark.

Through multi-threshold evaluation, learning curve reporting, ablation analysis, and honest failure assessment, we believe this work contributes toward a stronger transparency standard for medical AI research and lays groundwork for eventual clinical deployment, subject to the limitations above.

Author Contributions: Conceptualization, K.K. and H.M.R.; methodology, N.T.R., A.P. and H.M.R.; software, B.O.L.J. and A.P.; validation, H.M.R. and A.P.; resources, H.M.R., K.K., N.T.R. and B.O.L.J.; data curation, K.K., N.T.R. and A.P.; writing—original draft preparation, K.K., N.T.R. and B.O.L.J.; writing—review and editing, A.P. and H.M.R.; visualization, K.K. and A.P.; supervision, H.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this work is publicly available and may be freely downloaded from the given link: <https://www.kaggle.com/datasets/briscdataset/brisc2025/data>, accessed on 11 December 2025.

Acknowledgments: The part of the work carried out in the current article was supported by the project under the DST-SERB Category of EMEQ, Government of India, with File No. EEQ/2023/000053. The authors are thankful to the DST-SERB, Govt. of India, for supporting the work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fateh, A.; Rezvani, Y.; Moayedi, S.; Rezvani, S.; Fateh, F.; Fateh, M.; Abolghasemi, V. BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification. *Sci. Data* **2026**, *13*, 361. <https://doi.org/10.1038/s41597-026-06753-y>.
2. Chahal, P.K.; Pandey, S.; Goel, S. A survey on brain tumor detection techniques for MR images. *Multimed. Tools Appl.* **2020**, *79*, 21771–21814. <https://doi.org/10.1007/s11042-020-08898-3>.
3. Işın, A.; Direkoğlu, C.; Şah, M. Review of MRI-based brain tumor image segmentation using deep learning methods. *Procedia Comput. Sci.* **2016**, *102*, 317–324. <https://doi.org/10.1016/j.procs.2016.09.407>.
4. Pereira, S.; Pinto, A.; Alves, V.; Silva, C.A. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1240–1251. <https://doi.org/10.1109/tmi.2016.2538465>.
5. Banerjee, S.; Mitra, S.; Masulli, F.; Rovetta, S. Brain Tumor Detection and Classification from Multi-Sequence MRI: Study Using ConvNets. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 170–179. https://doi.org/10.1007/978-3-030-11723-8_17.
6. Amin, J.; Sharif, M.; Yasmin, M.; Fernandes, S.L. A distinctive approach in brain tumor detection and classification using MRI. *Pattern Recognit. Lett.* **2017**, *139*, 118–127. <https://doi.org/10.1016/j.patrec.2017.10.036>.
7. Amin, J.; Sharif, M.; Haldorai, A.; Yasmin, M.; Nayak, R.S. Brain tumor detection and classification using machine learning: A comprehensive survey. *Complex Intell. Syst.* **2021**, *8*, 3161–3183. <https://doi.org/10.1007/s40747-021-00563-y>.
8. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNET: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
9. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-UNet: UNET-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
10. Babar, N.A.; Lateef, J.; Syed, S.N.; Dietmeier, J.; O'Connor, N.E.; Raupp, G.B.; Spanias, A. Brain tumor classification in MRI scans using edge computing and a shallow Attention-Guided CNN. *Biomedicines* **2025**, *13*, 2571. <https://doi.org/10.3390/biomedicines13102571>.
11. Isensee, F.; Jäger, P.F.; Full, P.M.; Vollmuth, P.; Maier-Hein, K.H.. nnU-Net for Brain Tumor Segmentation. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 118–132. https://doi.org/10.1007/978-3-030-72087-2_11.
12. Khair, M.O.; Sharma, M.; Jain, V.; Chatterjee, J.M.; Jhanjhi, N.Z. A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. *IRBM* **2021**, *43*, 290–299. <https://doi.org/10.1016/j.irbm.2021.06.003>.
13. Xun, S.; Hang, Y.; Duan, S.; Chen, H.; Wang, M.; Chen, J.; Tan, T. RGA-UNET: An improved U-NET segmentation model based on residual grouped convolution and convolutional block attention module for brain tumor MRI image segmentation. In *Proceedings of the 2022 International Conference*; ACM: New York, NY, USA, 2022; pp. 319–324. <https://doi.org/10.1145/3569966.3570060>.
14. Gupta, R.K.; Bharti, S.; Kunhare, N.; Sahu, Y.; Pathik, N. Brain Tumor Detection and Classification Using Cycle Generative Adversarial Networks. *Interdiscip. Sci. Comput. Life Sci.* **2022**, *14*, 485–502. <https://doi.org/10.1007/s12539-022-00502-6>.
15. Gammoudi, I.; Ghazi, R.; Mahjoub, M.A. 3D Brain Tumor Segmentation Using Modified U-Net Architecture. In *Proceedings of the 2023 International Conference on Cyberworlds (CW)*; IEEE: Sousse, Tunisia, 2023; pp. 9–15. <https://doi.org/10.1109/CW58918.2023.00012>.
16. Xun, S.; Zhang, Y.; Duan, S.; Wang, M.; Chen, J.; Tong, T.; Gao, Q.; LAM, C.; Hu, M.; Tan, T. ARG-UNET: Advanced U-net segmentation model using residual grouped convolution and attention mechanism for brain tumor MRI image segmentation. *Virtual Real. Intell. Hardw.* **2024**, *6*, 203–216. <https://doi.org/10.1016/j.vrih.2023.05.001>.
17. Khaliki, M.Z.; Başarslan, M.S. Brain tumor detection from images and comparison with transfer learning methods and 3-layer CNN. *Sci. Rep.* **2024**, *14*, 2664. <https://doi.org/10.1038/s41598-024-52823-9>.
18. Aamir, M.; Namoun, A.; Munir, S.; Aljohani, N.; Alanazi, M.H.; Alsahafi, Y.; Alotibi, F. Brain tumor detection and classification using an optimized convolutional neural network. *Diagnostics* **2024**, *14*, 1714. <https://doi.org/10.3390/diagnostics14161714>.

19. Islam, M.M.; Rusho, M.A.; Ridita, N.N.; Shanta, Z.T.; Haider, S.; Jahan, A.F.H.D.K.; Qayum, M.A. Intelligent Systems in Neuroimaging: Pioneering AI techniques for brain tumor detection. *arXiv* **2025**, arXiv:2511.17655.
20. An, V.; Khajuria, A.; Pachauri, R.K.; Gupta, V. Multi-class classification of brain tumors using optimized CNN and transfer learning techniques. *Sci. Rep.* **2026**, *16*, 4709. <https://doi.org/10.1038/s41598-025-34806-6>.
21. Uniyal, M.; Saini, C.; Singh, D.P.; Ahmed, M.N.; Hussain, M.R.; Amarjeet, A.; Sinha, A.; Srivastava, A. Automated brain tumor detection using advanced deep learning models. *Discov. Artif. Intell.* **2026**, *6*, 1. <https://doi.org/10.1007/s44163-025-00753-4>.
22. Chatterjee, S.; Yassin, H.; Dubost, F.; Nürnberger, A.; Speck, O. Weakly-supervised segmentation using inherently-explainable classification models and their application to brain tumour classification. *Neurocomputing* **2026**, *682*, 133460. <https://doi.org/10.1016/j.neucom.2026.133460>.
23. Ke, L.; Hu, G.; Zhao, M.; Liu, Z.; Lv, Z.; Yang, Y. Brain tumor classification from MRI images using a multi-scale channel attention CNN integrated with SVM. *Sci. Rep.* **2026**, *16*, 6297. <https://doi.org/10.1038/s41598-026-36164-3>.
24. Srinivas, V.R.; Parvathi, R. Explainable AI-driven MRI-based brain tumor classification: A novel deep learning approach. *Front. Artif. Intell.* **2026**, *8*, 1700214. <https://doi.org/10.3389/frai.2025.1700214>.
25. Karian, Z.A.; Dudewicz, E.J. *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods*; CRC Press: Boca Raton, FL, USA, 2000.
26. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
27. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. *arXiv* **2015**, arXiv:1512.00567.
28. Vinogradova, K.; Dibrov, A.; Myers, G. Towards Interpretable Semantic Segmentation via Gradient-Weighted Class Activation Mapping (Student Abstract). *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13943–13944. <https://doi.org/10.1609/aaai.v34i10.7244>.
29. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. <https://doi.org/10.1007/s11263-019-01228-7>.
30. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Long Beach, CA, USA, 2017; pp. 4765–4774.
31. Yang, J.; Li, L.; Li, W.; Zheng, S. Harmonize task divergence with MUNet: Bridging semantic gaps between segmentation and classification for medical images. *Pattern Recognit.* **2026**, *180*, 114097. <https://doi.org/10.1016/j.patcog.2026.114097>.
32. Bajaber, A.S.; Ragab, M. Integrating whole-brain and masked-lesion CNN pipelines via multistage soft voting for enhanced three-class MRI tumor classification. *Biomed. Signal Process. Control* **2026**, *123*, 110594. <https://doi.org/10.1016/j.bspc.2026.110594>.
33. Pacal, I.; Deveci, M. The next generation of convolutional neural networks in neuro-oncology: A comparative analysis of performance, model complexity, and efficiency in brain tumor diagnosis. *Brain Netw. Disord.* **2026**, *2*, 136–146. <https://doi.org/10.1016/j.bnd.2026.02.001>.
34. Ahmed, F. Enhancing brain tumor classification using Vision Transformers with colormap-based feature representation on BRISC2025 dataset. *arXiv* **2026**, arXiv:2603.21234. <https://doi.org/10.48550/arXiv.2603.21234>.
35. Thahiruddin, M.; Wulandari, A. CNNs vs. Hybrid Transformers for brain tumor classification on the BRISC dataset. *J. Apl. Teknol. Inf. Dan Manaj. (JATIM)* **2025**, *6*, 24–33. <https://doi.org/10.31102/jatim.v6i1.3545>.
36. Prayogo, R.D.; Karimah, S.A.; Nambo, H. HayabusaNet: Hybrid attention-based multiscale fusion CNN for accurate and efficient brain tumor classification in MRI scans. *Expert Syst. Appl.* **2026**, *331*, 133135. <https://doi.org/10.1016/j.eswa.2026.133135>.
37. Addai, G.E.; Zhang, X.; Ukwuoma, C.C. Content-adaptive multi-scale graph attention with lesion-focused pooling for multi-class brain tumor diagnosis. *Appl. Soft Comput.* **2026**, *196*, 115100. <https://doi.org/10.1016/j.asoc.2026.115100>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.