Journal of Theoretical and Applied Information Technology
15th September 2025. Vol.103. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NEUROEXPLAINAI: AN EXPLAINABLE AI AND STATISTICAL FRAMEWORK FOR BRAIN TUMOR DIAGNOSIS AND SEVERITY PREDICTION USING MULTIMODAL MRI WITH NEUROFUSIONNET

V.HARI PRASAD[1*], DR. T. BHASKAR[2], SUKANYA LEDALLA[3], M. VARAPRASAD RAO[4], A. HARSHAVARDHAN[5]

[1]Lecturer in Computer Engineering, Govt.Polytechnic PRODDATUR, Department of Technical Education ,Mangalagiri Andhra Pradesh, India.
[2]Associate professor, Department of Computer Science and Engineering (AI & ML), CMR College of Engineering andTechnology, Kandlakoya, Medchal, Hyderabad -501401.
[3]Assistant Professor Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad-500043, Telangana, India.
[4]Professor, Department of CSE(DS), CVR College of Engineering, Hyderabad, Telangana,India.
[5]Sr. Asst. Prof, Dept. of CSE (AIML&IoT), VNR Vignana Jyothi Institute ofEngineering and Technology, Hyderabad, Telangana, India.
hariprasadvemulapati@gmail.com, bhalu7cs@gmail.com, ledalla.sukanya@gmail.com, varam78@gmail.com, harshavgse@gmail.com
*Corresponding Author

ABSTRACT

The correct diagnosis and prediction of malignancy in brain tumors are critical for neuro-oncology, as they directly influence clinical decision-making. Although deep learning models have had notable success in tumor classification and segmentation based on MRI data, most existing approaches are limited in three aspects: building on imaging modalities only, disregarding clinically relevant metadata, and lacking interpretability because of non-integrated explainable AI (XAI). To overcome these limitations, we present NeuroExplainAI, an explainable deep learning framework for a holistic brain tumor diagnosis and grading. We present NeuroFusionNet, a dual task architecture to fuse deep CNN features with hand crafted radiomic descriptors and patient-level clinical metadata in the form of data-attention. This allows for classification (HGG versus LGG) and severity scoring to be performed concurrently. For decisor transparency, both spatial and channel-level explanations are included using Grad CAM++ and SHAP. The model is trained and tested on BraTS 2021 dataset with 98.34% accuracy, 97.73% F1-score and MAE=0.38 for severity prediction. This paper provides novel insights into the clinical interpretability of multimodal fusion and attention-based weighting, in addition to its effect on the predictive performance. Ablation study and comparisons with state-of-the-arts demonstrate the necessity and effectiveness of each component. The incorporation of explainableAI techniques builds trust and improves usability in clinical workflows, making NeuroExplainAI an appealing platform for reliable, interpretable, and individualized brain tumor assessment.

Keywords - *Brain Tumor Diagnosis, Multimodal MRI, Explainable AI, Severity Prediction, Deep Learning Framework*

1. INTRODUCTION

Brain tumors are among the most critical and life-threatening conditions in neurology, often requiring rapid and accurate diagnosis to guide treatment strategies. Magnetic Resonance Imaging (MRI) is central toidentifying and assessing brain tumors due to its high spatial resolution and multiparametric capability. However, manual interpretation of MRI scans is subject to inter-observer variability, and traditional radiological assessments may fail to capture subtle imaging cues. Deep learning techniques have emerged as powerful tools for automating brain tumor diagnosis and segmentation to address these challenges, offering high accuracy and consistency. While several recent models [1], [2] have demonstrated

success in tumor classification or segmentation, most rely solely on imaging data, neglecting valuable non-imaging information such as clinical metadata. Moreover, the lack of explainability in many existing deep learning models poses a significant barrier to clinical adoption.

Recent studies have explored hybrid models combining CNNs with vision transformers [2], ensemble learning [3], and attention mechanisms [4], showing improvements in classification and segmentation tasks. Nevertheless, few approaches support dual-task learning for classification and severity prediction, and even fewer integrate radiomic features and clinical metadata into the decision-making pipeline. There remains a critical need for interpretable, multimodal AI systems capable of providing comprehensive diagnostic outputs that clinicians can trust and validate.

This research proposes NeuroExplainAI, an explainable and integrated AI framework for brain tumor diagnosis and severity prediction using multimodal MRI to address these gaps. The primary objective is to design a deep-learning model that classifies tumor types (HGG vs. LGG) and predicts tumor severity scores using a dual-task architecture. The proposed system introduces several key novelties: (1) multimodal feature fusion of deep CNN features, radiomic descriptors, and clinical metadata; (2) an attention-guided fusion layer to prioritize informative features; and (3) explainability via Grad-CAM++ and SHAP to ensure transparency in decision-making. Additionally, the framework includes statistical analysis to correlate tumor features with severity outcomes, enhancing clinical insight.

The contributions of this work are multifold. First, we develop a robust NeuroFusionNet architecture that performs classification and regression jointly. Second, we demonstrate the impact of radiomics and clinical metadata in improving diagnostic performance. Third, we implement a dual-level explainability module that visualizes spatial and feature-level attributions. We conduct extensive evaluations, including ablation studies and comparisons with state-of-the-art methods.

Notwithstanding the increasing enthusiasm in the field of deep learning for brain tumor diagnosis, most prior frameworks are confined either to a single-task learning or a particularly narrow application on imaging data, largely dismissing a wealth of complementary information extracted from radiomic features and patient-specific clinical metadata. Additionally, the black-box nature of these models is a major obstacle to the clinical acceptance of these models because clinicians cannot interpret the outputs of the model or validate the decisions of the model. Since brain tumors present a wide range of morphological and contextual patterns from patient to patient, an overall interpretable and unified system, which integrates multimodal information, is in high demand. This study fills these key gaps by presenting a holistic interpretable and dual-task AI model -- NeuroExplainAI -- aiming at both accurate tumor classification, severity scoring and providing a transparency to the decision-making. The use of spatial (Grad-CAM++) and attribution-based (SHAP) explainability modules strengthens clinical trust, distinguishing this work from classic black-box models, and sets a new state-of-the-art for diagnostic assistance in neuro-oncology.

The remainder of this paper is organized as follows: Section 2 provides an in-depth literature review on explainable deep learning-based brain tumor diagnosis. Section 3 details the proposed NeuroExplainAI methodology, which includes the preprocessing stage, segmentation process, feature extraction and description, model architecture, and explainable AI integration. Section 4 presents the experimental results, including performance analysis, ablation studies, and comparative evaluations. Section 5 discusses the findings and limitations of the study in detail. Finally, Section 6 summarizes the article with some takeaways and directions for future research.

2.RELATED WORK

This literature review highlights recent trends in explainable AI and deep learning techniques for braintumor diagnosis based on multi-modal MRI. Zeineldin et al. For example, [1] examined the incorporation of explainability into deep neural networks used for MRI-based brain tumor analysis, facilitating greater interpretability for clinical decision-making. Zeineldinet al. [2] recent work, where they employed vision transformers along with CNNs in a hybrid approach of multimodal glioma segmentation that performed with high accuracies and explainable study outputs. Farhan et al. [3] introduced an ensemble 3D brain tumor segmentation technique in XAI-MRI using dual-

modality MRIs with attention-based explainability. Ahmed et al. Proposed a hybrid ViT-GRU model for brain tumor classification in Bangladesh, with explanations through XAI visualization to improve the transparency of the model. Maqsood et al. [5] employed deep neural networks and support vector machines (SVM) for multimodal tumor detection, achieving promising classification performance. Aleid et al. AI-based MRI analysis was used by [6] to deploy an early detection method. Di Noia et al. [7] discussed AI approaches used in outcome prediction, specifically in MRI-based prognostics. Anand et al. [8] present a multimodal segmentation classification pipeline with machine learning integration. Gesperger et al. His work [9] applied deep learning and multimodal microscopy to enhance diagnostic imaging. Khalighi et al. [10] explored the state of AI in neuro-oncology, emphasizing diagnosis, prognosis, and precision therapy.

Li et al. [11] introduced a deep learning framework for hemorrhagic lesion detection and segmentation in brain CTs, showcasing transferability to tumor detection tasks. Amin et al. [12] developed a CNN-based brain tumor classification model using MRI scans, emphasizing high-speed detection. Saba et al. [13] proposed a hybrid model that fuses handcrafted and deep features, improving classification accuracy and robustness. Amin et al. [14] leveraged stacked autoencoders for automatic tumor detection, offering a hierarchical representation of tumor features. Oksuz [15] addressed MRI artifacts using CNNs, indirectly improving preprocessing for brain tumor analysis. Woźniak et al. [16] used a correlation learning mechanism for CT-based tumor detection, highlighting potential cross-modality generalizability. Kalaiselvi et al. [17] utilized pseudo coloring to enhance multimodal MRI features for tumor detection. Using multimodal MRI, Sun et al. [18] developed a deep-learning pipeline for tumor segmentation and survival prediction. Li et al. [19] proposed a CNN-based model integrating multimodal fusion for tumor classification. Peng and Sun [20] introduced AD-Net for multimodal segmentation, achieving high performance using attention-driven fusion.

Kermi et al. [21] applied a U-Net-based deep CNN for brain tumor segmentation using multimodal MRI, achieving substantial spatial accuracy. Windisch et al. [22] emphasized explainability by integrating model interpretability into CNNs for essentialtumor detection using MRI slices. Atasever et al. [23] provided a comprehensive survey on medical image analysis using deep learning, focusing on the significance of transfer learning in diagnostic tasks. Tripathy et al. [24] implemented EfficientNet for brain tumor classification from MRI, improving both accuracy and computational efficiency. Preetha et al. [25] conducted a comparative study of deep neural network architectures for tumor segmentation, evaluating performance across different backbones. Ahmad and Choudhury [26] assessed transfer learning models for brain tumor detection, highlighting VGG and ResNet as strong performers. Anaya-Isaza and Jiménez [27] used data augmentation with transfer learning to enhance classification from MRI. Khan et al. [28] proposed a deep CNN framework with high tumor detection accuracy. Solanki et al. [29] reviewed intelligent techniques for tumor classification. Ottom et al. [30] introduced ZNet for 2D tumor segmentation with improved boundary delineation.

Table 1: Literature Review Summary Of Selected Related Works On Brain Tumor Diagnosis Using Deep Learning

Author and Year	Model / Technique	Modality	Task	Explainability	Key Limitations
Zeineldin et al. [2]	Hybrid ViT + CNN	Multimodal MRI	Segmentation	Grad-CAM	No severity prediction; lacks clinical metadata integration
Farhan et al. [3]	Ensemble CNNs	T1, T2 MRI	3D Segmentation	SHAP, Grad-CAM	No classification; lacks metadata and global interpretability
Ahmed et al. [4]	ViT + GRU	Multimodal MRI	Classification + Risk	Attention-based XAI	No radionics; lacks spatial explainability

Maqsood et al. [5]	DNN + Multiclass SVM	Multimodal MRI	Detection	Not mentioned	No explainability; limited field adaptability
Amin et al. [12]	CNN Classifier	MRI	Classification	Not included	Single modality; lacks severity scoring
Kermi et al. [21]	U-Net (CNN)	Multimodal MRI Volumes	Segmentation	Not included	Segmentation only; lacks explainability
Haque et al. [35]	NeuroNet19 (DNN)	MRI	Classification	Layer-wise Relevance	No severity prediction; lacks multimodal integration
Hosny et al. [36]	Ensemble Deep CNN	T1, T2, FLAIR	Detection + Grading	Grad-CAM++, SHAP	No attention fusion; limited clinical metadata use
Sun et al. [18]	Deep CNN	Multimodal MRI	Segmentation + Survival	Not clearly stated	No classification; lacks interpretability
Hassan et al. [38]	XAI-CNN	MRI	Segmentation	SHAP	No classification; no metadata integration

Talukder et al. [31] proposed a fine-tuned deep-learning model integrating reconstruction mechanisms for improved tumor categorization using MRIs. Anaya-Isaza et al. [32] presented a comparative analysis of neural architectures for MRI-based brain tumor detection, including cross-transformers and transfer learning. Nhlapho et al. [33] focused on bridging the interpretability gap in deep models, offering insights into explainable AI for MRI diagnosis. Taşcı [34] introduced DGXAINet, which integrates attention-based deep feature extraction with explainable learning for tumor localization. Haque et al. [35] proposed NeuroNet19, an explainable DNN architecture for brain tumor classification with high interpretability. Hosny et al. [36] developed an explainable ensemble model using multiple deep learners for detection and grading. Sinha et al. [37] introduced an XAI-enhanced model that aids clinicians in tumor assessment. Hassan et al. [38] unfolded the structure of explainable models for accurate tumor segmentation. Li and Dib [39] emphasized trustable diagnosis using explainable deep learning. Naira et al. [40] built an explainable diagnostic model utilizing discharge summaries for MRI-based tumor classification.Table 1 summarizes key literature on brain tumor diagnosis, comparing models' tasks, modalities, and limitations. NeuroExplainAI outperforms existing methods by integrating multimodal data and explainability. The reviewed studies demonstrate a growing emphasis on explainability, multimodal fusion, and hybrid architectures in brain tumor analysis. Techniques span CNNs, transformers, radionics, and transfer learning, with several works integrating saliency maps or SHAP for interpretability. They highlight the importance of accuracy, clinical trust, and robust multimodal diagnostic systems.

The literature review presents many improvements in brain tumor grading and segmentation via deep learning. But a closer look reveals some long overdue deficiencies, albeit ones that are largely unaddressed. For instance, Zeineldin et al. [2] and Farhan et al. [3] implemented hybrids and ensembles which performed well in segmentation but did not support severity scoring or use of patient-level metadata. Ahmed et al. [4] performed XAI-driven classification, but did not include hand-crafted radiomic features or validated predictions using statistical analysis. In the same way, the models like Maqsood et al. [5] and Kermi et al. [21] only considered imaging data without explainability or multimodal fusion. However, these studies, technically remarkable as they are, either emphasize performance over explainability, or lack overall diagnostic insights. Third, the lack of frameworks that incorporate dual-task learning (classification + severity), multi-source feature integration and explainable

AI modules also exists a significant disparity in the practical application of these methods for real-life clinical practice. This problem space represents an unreasonable demand for a holistic, interpretable, and multimodal framework, which NeuroExplainAI specifically addresses and intends to ameliorate its shortcomings to improve both diagnostic accuracy and clinical utility.

3. PROPOSED FRAMEWORK

The proposed framework, named NeuroExplainAI, is a novel, explainable AI and statistical framework developed for automated diagnosis and severity prediction of brain tumors using multimodal MRI data. The system integrates deep learning, radiomic feature analysis, and clinical metadata to enhance prediction accuracy and clinical interpretability. It leverages a dual-path architecture for tumor classification (HGG vs. LGG) and severity scoring (e.g., WHO grade), supported by a comprehensive XAI module using Grad-CAM++ and SHAP for image-space and feature-space explanations, respectively. The core deep learning model, NeuroFusionNet, fuses modality-specific CNN features, handcrafted radiomic descriptors, and patient metadata through an attention-guided fusion layer. Additionally, statistical analysis complements the model outputs with volumetric and correlation-based insights to improve clinical relevance.

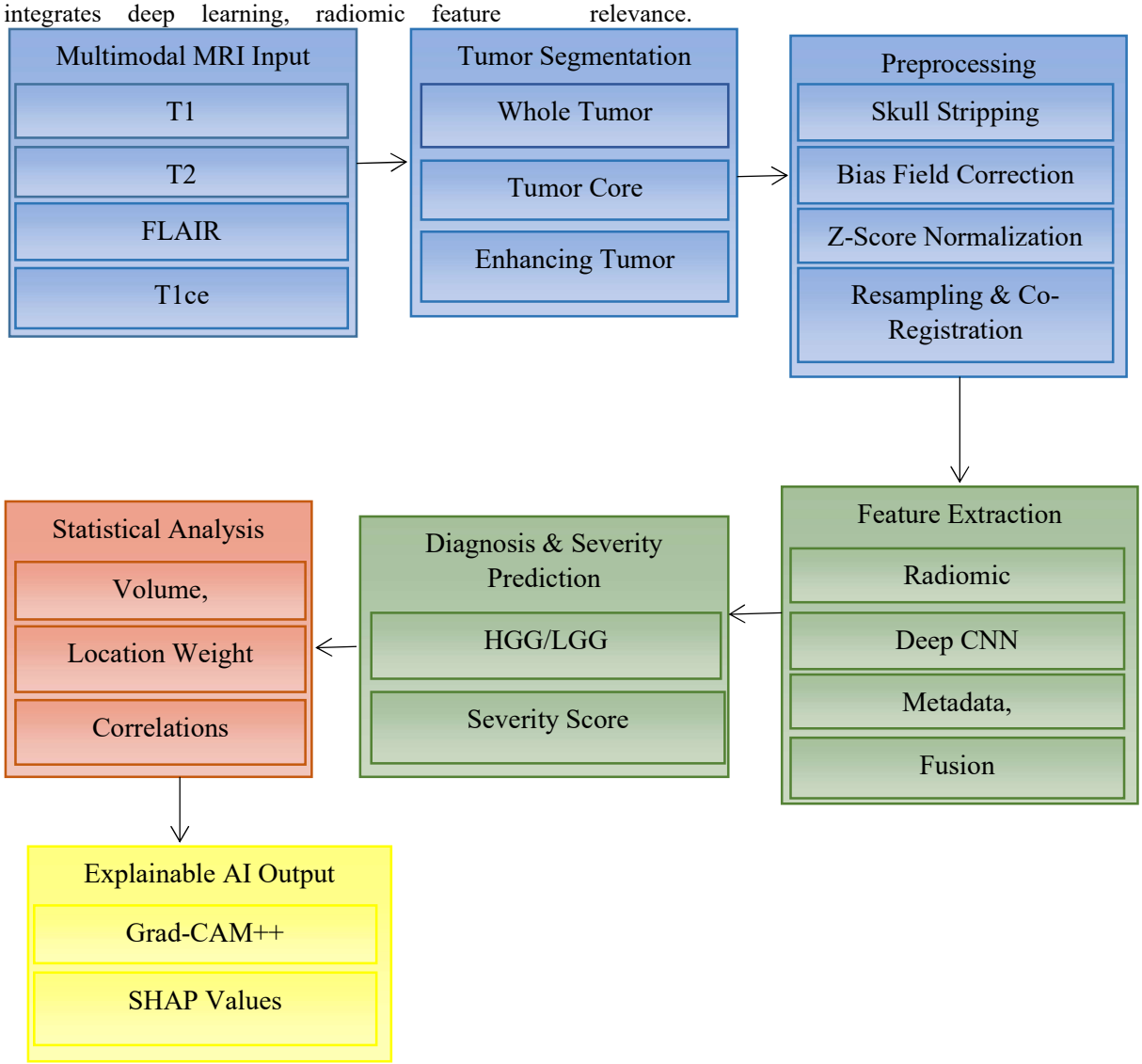


Figure 1: System Architecture Of Neuroexplainai Illustrating The Complete Pipeline For Brain Tumor Diagnosis And Severity Prediction

Figure 1 The NeuroExplainAI framework uses multimodal MRI data to automate brain tumor diagnosis and severity prediction. The system begins with the input of four MRI modalities—T1, T1ce, T2, and FLAIR—which undergo standardized preprocessing steps, including skull stripping, bias field correction, z-score normalization, and spatial resampling with co-registration. Segmentation of tumor subregions, such as the whole tumor and tumor core, and enhancing tumors is carried out using an advanced segmentation model, facilitating precise region-of-interest extraction.

Subsequently, features are extracted from multiple sources: radiomic features from segmented regions, deep features from CNN encoders applied to MRI inputs and clinical metadata such as patient age and gender. These features are fused into a unified representation and passed to the NeuroFusionNet model, which performs dual tasks: classifying tumor type as HGG or LGG and predicting severity scores. Statistical analysis involving tumor volume, location weighting, and correlation metrics enhances interpretability, while Grad-CAM++ and SHAP explainability outputs offer visual and feature-level insights. Table 2 presents the notations used in the proposed methodology.

Table 2: Notations used in the methodology

Symbol (s)	Description
$X \in \mathbb{R}^{4\times H\times W\times D}$	Multimodal MRI input volume (T1, T1ce, T2, FLAIR)
$V \in \mathbb{R}^{H\times W\times D}, V'$	Single MRI modality volume and its z-score normalized version
$\mu v, \sigma v$	Mean and standard deviation of non-zero voxels in V
$M \in \{0,1\}^{H\times W\times D}$	Binary segmentation mask for tumor subregions (ET, TC, WT)
$f_{CNN}^{(i)}$	CNN feature extractor for modality ii
F_d, F_r, F_c	Deep, radio mic, and clinical metadata feature vectors, respectively
F, F', F''	Concatenated, fused, and attention-weighted feature vectors
α	Attention weights for recalibrating fused features
$\hat{y}_{class}, \quad \hat{y}_{sev}$	Predicted tumor class (HGG/LGG) and severity score or grade
y_k, y_{sev}	Ground truth class label and severity score
$W_f, b_f, W_{class}, b_{class}, W_{sev}, b_{sev}$	Weights and biases in fusion, classification, and regression layers, respectively
$Phi\phi(.)\ \sigma(.)$	ReLU and sigmoid activation functions
$\odot, \Vert$	Element-wise multiplication and vector concatenation operators
A^k, α_k^c	CNN feature map and its importance weight in Grad-CAM++
$L_{Grad-CAM++}^c$	Saliency map highlighting discriminative regions for class c
ϕ_0, ϕ_i	SHAP base value and feature attribution for x_i
S	Location-weighted severity score
T_X, T_Y, T_Z	Voxel resolutions along each axis (used in volume computation)
$f(x)$	Model output decomposed by SHAP into contributions.
$s(.)$	Attention or scoring function used for relevance computation

3.1Data Acquisition and Preprocessing

This study is based on the open-access BraTS 2021 dataset [41], which consists of pre-operative multimodal MRI of patients with brain tumors, targeting four imaging modalities: native T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and fluid-attenuated-inversion-recovery (FLAIR) imaging. Data from each subject contains co-registered, skull-stripped, and resampled volumes with a standardized voxel size of 1 mm³ (isotropic), normalizing the practical spatial dimension of all inputs. The dataset contains expert annotated ground truth regions of interest for tumor subregions, including enhancing tumor (ET), tumor core (TC), and whole tumor (WT), along with corresponding tumor grade labels (high-grade glioma (HGG), low-grade glioma (LGG)).

Although the dataset was preprocessed, normalization across subjects was performed to prevent the abovementioned artifacts from troubling the data. This was the z-score normalized for each modality volume.$V \in \mathbb{R}^{H\times W\times D}$ in the Eq. 1.

$$V' = \frac{V - \mu V}{\sigma V} \qquad (1)$$

Where μV and μV Are the mean and standard deviation of the non-zero voxels in the volume, respectively? Standardized all intensity distributions are to be equivalent across patients to help reduce scanner and subject variability. A 3D reslicing operation standardized all MRI volumes, $240 \times 240 \times 155$ Allowing batch input for the neural network pipeline.In the preprocessing step, skull stripping was again confirmed by applying a binary brain mask to remove any remaining extraneous brain tissue. Also, the segmentation masks for tumor subregions were encoded into three separate channels, with each channel corresponding to each tumor region (ET, TC, WT). They were applied to monitor both the segmentation phase and the extraction of radiomic characteristics in anatomically relevant areas of the following phases. The preprocessed dataset retains spatial and intensity homogeneity, providing a solid base for downstream deep learning and statistical analysis.

3.2Tumor Segmentation

The authors use a variation of the U-Net++ architecture, which is modified to extract local and global spatial features from the multimodal MRI inputs for tumor segmentation. Each patient's input is a four-channel 3D volume.$X \in \mathbb{R}^{4\times H\times W\times D}$where the four channels correspond to modalities T1, T1ce, T2, and FLAIR. The segmentation task learns a mapping functionf_θ That predicts voxel-wise predictions.$\hat{Y} = f_\theta(X)$with $\hat{Y} \in \mathbb{R}^{C\times H\times WD}$ and $C = 3$ The following represent tumor subregions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT), respectively.

The classical encoder-decoder structure has been extended to a modified U-Net++ architecture, including attention gates and SE blocks. The attention gates enhance the model's focus on tumor regions by inhibiting irrelevant background activations, while SE blocks adaptively recalibrate channel-wise feature responses, facilitating informative feature representations. The encoder comprises several convolution blocks containing 3D convolution layers followed by batch normalization and ReLU activation. 3D max-pooling layers are used for downsampling, and 3D transposed convolutions are used for upsampling during the decoder path. Skipping connectivity between the encoding and decoding layers is retained and is dense to preserve spatial details and prevent the loss of features.

The model is trained with a compound loss function.$\mathcal{L}_{seg}$ Whichdescribes the sum of the Dice loss and the binary cross entropy (BCE) loss, expressed in Eq. 2.

$$\mathcal{L}_{seg} = \lambda_1 . \mathcal{L}_{Dice} + \lambda_2 . \mathcal{L}_{BCE} \quad (2)$$

Where λ_1 and λ_2Are weighting parameters empirically adjusted to balance the overlap accuracy and voxel-wise precision. For managing class imbalance, the Dice loss not only maintains overlap with groot truth tumor regions but also ensures that the segmentation model has the highest overlap with the ground truth.

In the first step, segmentation output gives multi-label masks for each tumor subregion that the radiomic feature extraction module will use as input; in the second stage, they are used as auxiliary information to visualize tumor structures (explainable AI) in the final stage. Segmentation module of NeuroExplainAI pipeline Anatomically accurate segmentation provides a fundamental first step for subsequent downstream analysis.

3.3Feature Extraction

After tumor segmentation, a full-feature extraction method is utilized to achieve multi-domain representations in diagnosis and prognosis prediction of tumor severity. Depending on their interpretation, the features are grouped into three categories: the deep features, the radiomic features, and the clinical metadata. In this sense, we expect the multi-source feature strategy to enrich and generalize the representation of the downstream NeuroFusionNet model.

Dal city -$M_i \in \{T1, T1ce, T2, FLAIR\}$specific MRI volume passes through separate 3D convolutional branches to learn deep features—a compact 3D CN encoder ed to learn each modality's hierarchical spatial features in brain tissue. We flatten and concoutputs from these modality-specific ranches to generate a unified dep feature vector, cap F sub d, element of double-double-struck cap component n sub d, end superscript, $F_d \in \mathbb{R}^{n_d} n_d$is the dimensionality of the learned features. These features capture high-level semantic information, including tumor location, shape, and texture across modalities.

PyRadiomics toolkit is used to extract radiomic features from the segmented tumor regions. A feature set, which includes first-order statistics, shape descriptors, and texture features calculated using the Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), and other related approaches, is

computed from each subregion, i.e., enhancing tumor (ET), tumor core (TC), and whole tumor (WT). We denote the resultingradio mic feature vector as $F_r \in \mathbb{R}^{n_r}$ Such features provide quantitative information of tumor heterogeneity and morphology that supplements what is captured in these deep learning-based model representations."

Normalized numerical vector $F_c \in \mathbb{R}^{n_c}$To encode clinical metadata such as patient age, gender, and tumor location (if available). Except for the targets, all non-numerical variables are one-hot encoded, and continuous variables are min-max scaled to the range [0, 1] so that neural network input is compatible.All three sources are concatenated to form the final composite feature vector. $F \in \mathbb{R}^n$ as in Eq. 3.

$$F = [F_d \| F_r \| F_c] \quad (3)$$

where $\|$ is vector concatenation, and $n = n_d + n_r + n_c$This merged feature vector is the input to the classification and regression branches of NeuroFusionNet architecture. The feature extraction module guarantees that sufficient tumor information is extracted for subsequent prediction tasks by incorporating complementary information and utilizing learned, handcrafted, and clinical features from all domains.

3.4 Feature Fusion Layer

The resulting feature vector for deep, radiomic, and clinical domains is fused into a final output for input to a specialized feature fusion layer to improve joint representation learning and reduce the effects of potential modality imbalance. Such a fusion layer is an integrative bridge across heterogeneous features by mapping them into a unified latent space for classification and severity regression tasks. Fusion is realized in multiple fully connected layers with non-linear activations, dropout regularization, and an additional optional attention-based weighting strategy.Given the concatenate feature sector cap F element of double-struck cap R to the tor cap F element of double-struck cap R to the tor $F \in \mathbb{R}^n$ Defined in the preceding paragraph, the first transformation stage is a fully connected (FC) projection defined as in Eq. 4.

$$F' = \phi(W_f F + b_f) \quad (4)$$

Where $W_f \in \mathbb{R}^{n' \times n}$ Is the learnable weight matrix, $b_f \in \mathbb{R}^{n'}$ is the biased term and $\phi(.)$ is the ReLU activation function. This output $F' \in \mathbb{R}^{n'}$Transition is the latent fusion of feature embedding.

During the training phase, a dropout layer with dropout_rate$p = 0.3$ (to improve robustness in low-frequency interpretation and reduce over-fitting) is added. In addition, we develop an optional self-attention mechanism for adaptive anatomy of the importance of each type of feature. This mechanism instead calculates a set of attention weights.$\alpha \in \mathbb{R}^{n'}$ via a learnable scoring function $s(.)$ and applies them in an element-wise fashion to F' as in Eq. 5.

$$F'' = \alpha \odot F' \text{where} \quad \alpha = \sigma(s(F')) \quad (5)$$

where, $\odot$ is element-wise multiplication and, $\sigma(.)$ is the sigmoid activation function. The output $F'' \in \mathbb{R}^{n'}$ It is the final fused representation that incorporates cross-feature dependencies and relevance scaling.

The resulting fused vector is passed on parallel to a classification head to predict the type of brain tumor (HGG vs LGG) and a regression head to predict severity (WHO grade or risk score). The feature fusion layer hence acts as the hub to harmonize information from the multimodal, handcrafted, and clinical domains, allowing NeuroFusionNet to learn a rich and interpretable representation space.

3.5 NeuroFusionNet Architecture

We present the NeuroFusionNet architecture, a hybrid deep learning model catering to multi-modality segmentation input to provide dual outputs: brain tumor classification and severity grading. It employs parallel convolutional encoders for the MRI modalities, combines handcrafted radio mic and clinical features, and aggregates them through a shared fusion layer. The architecture comprises three functional components: modality-specific feature encoders, a fusion and transformation core, and dual-task output branches.

Journal of Theoretical and Applied Information Technology
15th September 2025. Vol.103. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

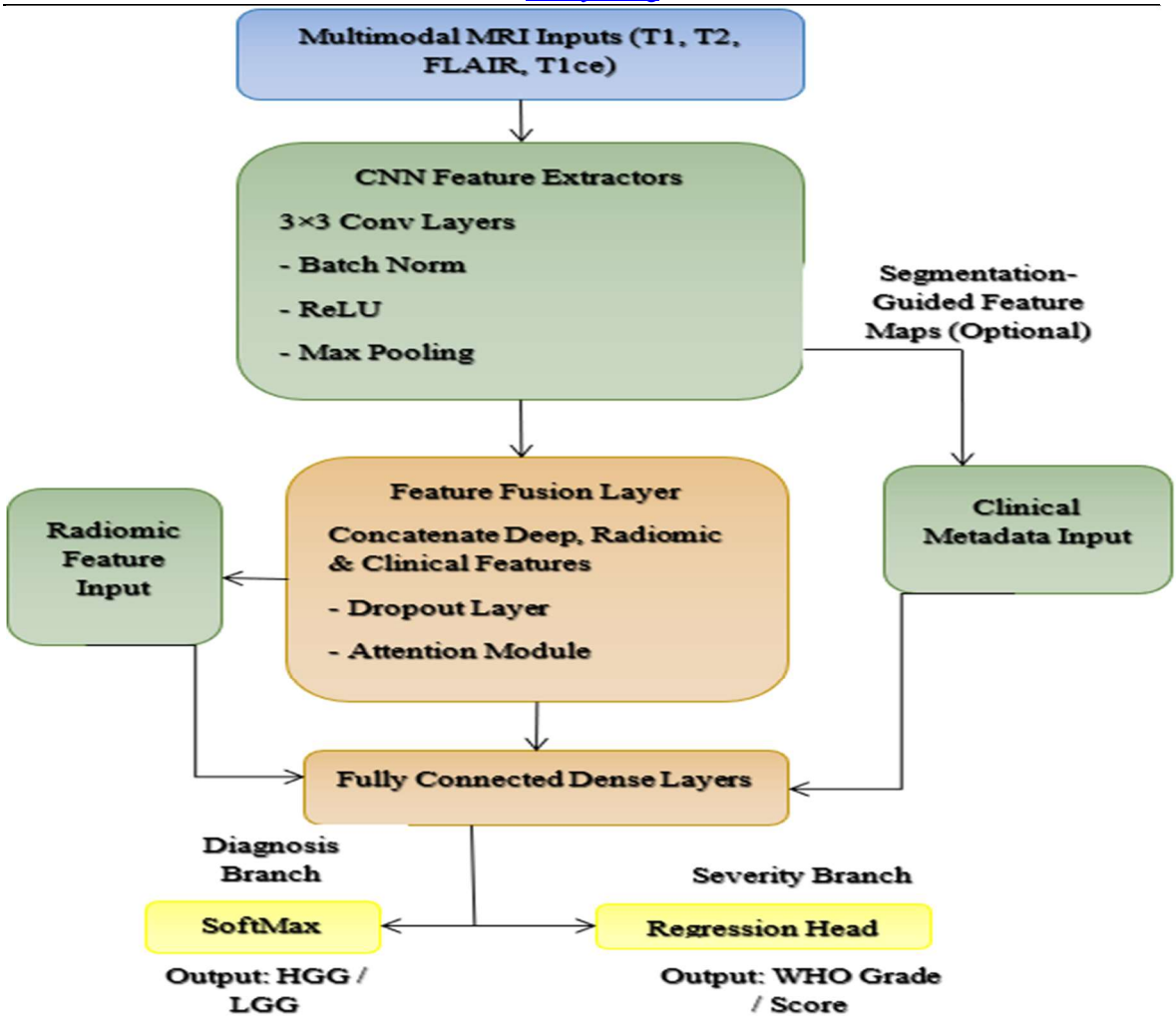


Figure 2: Model Architecture Of Neurofusionnet For Brain Tumor Diagnosis And Severity Prediction

Figure 2 To extract its respective deep features, a dedicated 3D convolutionalencoder$f_{CNN}^{(i)}$is used for each modality — T1, T1ce, T2, and FLAIR. A stacked version of 3D convolution layers, batch normalization, ReLU activations, and max pooling operations form an encoder. All four encoder outputs are flattened and concatenated into single deep to cap F sub d , meanwhile to cap F sub d , meanwhile to cap F sub d , sub d , meanwhile to cap F sub d , meanwhile—meanwhile.$F_d \in \mathbb{R}^{n_d}$. Meanwhile, hand-crafted radiomic features F_r ∈ted and prepared separately as inputs. The joint representation $F \in \mathbb{R}^n$ of these three feature vectors is achieved according to the fusion strategy defined above.

The fused feature vector $F'' \in \mathbb{R}^{n'}$ It is then passed through a shared transformation layer consisting of fully connected layers, dropout, and nonlinear activations. The standard layer allows for more task-agnostic representation learning before branching into two separate output heads—a classification and regression head.

This is a classification branch meant for predicting tumor type (HGG vs LGG). It consists of a fully connected layer with a softmax activation function. The predicted class probabilities are in Eq. 6.

$$\hat{y}_{class} = softmax(W_{class}F'' + b_{class}) \quad (6)$$

Where $W_{class}$and b_{class} are learnable parameters and. The loss function applied is categorical cross-entropy, given by Eq. 7.

$$\mathcal{L}_{class} = -\sum_{k-1}^{K} y_k \log(\hat{y}_k) \quad (7)$$

The true label y_k is one-hot encoded where $K = 2$ is used for binary classification (HGG/LGG), and the severity prediction branch works in regression mode to create a continuous severity score or WHO grade. It employs a dense output

layer with a linear activation, resulting in the prediction in Eq. 8.

$$\hat{y}_{sev} = W_{sev}F'' + b_{sev} \quad (8)$$

Where W_{sev} and b_{sev}Are regression weight and bias, respectively. The loss in severity is represented using mean squared error (MSE), as in Eq. 9.

$$\mathcal{L}_{sev} = \frac{1}{N}\sum_{i-1}^{N}\left(\hat{y}_{sev}^{(i)} - y_{sev}^{(i)}\right)^2 \quad (9)$$

Eq. 10 gives the overall training objective, a unified loss function that combines classification and regression losses.

$$\mathcal{L}_{total} = \lambda_1, \mathcal{L}_{class} + \lambda_2 . \mathcal{L}_{sev} \quad (10)$$

where λ_1 and λ_2 Are the balancing hyperparameters. This dual-output architecture allows NeuroFusionNet to conduct multi-task learning, improving the overhead of the generalization ability of the model while simultaneously offering diagnostic and prognostic information.

3.6Statistical Analysis

NeuroExplainAI has a deep learning-based prediction module that is added to a statistical analysis module to ensure interpretability from a clinical perspective and complement profound learning-based predictions. Module 1: Quantitative measurements extracted from segmented tumor regions and their association with tumor severity and clinical metadata. It can be used to determine the statistical significance of morphological and spatial features related to predicted tumor grade/severity scores.

The first stage is to calculate volumetric metrics for each subregion of the tumor, specifically the enhancing tumor (ET), tumor core (TC), and whole tumor (WT). The corresponding volume V in cubic millimeters is calculated from $M \in \{0,1\}^{H\times W\times D}$ A binary segmentation mask by Eq. 11.

$$V = \sum_{i-1}^{H}\sum_{j-1}^{W}\sum_{k-1}^{D} M_{i,j,k} . r_x . r_y . r_z \quad (11)$$

Where $r_x . r_y . r_z$Are the voxel sizes (in mm) along each dimension. They are critical components in assessingtumor burden and progression.Apart from some volume, we also derived spatial features, including tumor location, by calculating each tumor region's center of mass (CoM). Tumor anatomical location is mapped to the Center of Mass (CoM). It is projected onto brain region atlases, which are then assigned specific weights based on tumor proximity to clinically critical domains. We define a location-weighted severity metric.S as in Eq. 12.

$$S = \sum_{i-1}^{n} w_i . f_i \quad (12)$$

Where f_i denotes a region-specific morphology or intensity feature and w_iIs the location-based risk weight.Correlation and hypothesis testing are used to assess relationships between these quantitative features and the severity of the tumor. To calculate linear associations between tumor volume and severity score, we use r Pearson correlation coefficient. Spearman's rank correlation is applied when normality assumptions are violated. For hypothesis testing, feature distributions between HGG and LGG classes are compared using two-sample t-tests. Also, we used one-way ANOVA to analyze multi-group differences by stratifying based on the WHO grades.

P-value thresholds $p < 0.05$ Are reported for statistical significance. All analyses are performed using standard Python libraries like SciPy and StatsModels. These statistical patterns confirm NeuroFusionNet's predictions and uncover interpretable relationships between anatomical features and disease severity, adding to clinical relevance.

3.7Explainable AI Integration

This is possible by incorporating a dual-stage explainable AI (XAI) module in the NeuroExplainAI framework that centerson interpretability in image space and attribution in the feature space to maintain the transparency and clinical trust of the clinical predictions generated by the NeuroExplainAI framework. The detailed insights from the module are expected to enable clinicians and researchers to comprehend the months' concepts used by the NeuroFusionNet model for decision formations, enabling its integration in clinical decision-making situations.

In the image domain, we use Grad-CAM++ on the last convolutional layers from the CNN branches, which process each MRI modality. This method creates class-discriminative saliency maps highlighting the spatial regions with the most response frequency, contributing to the model prediction. To have a predicted class score y^c Grad-CAM++ calculates the weight.α_k^c for each feature mapA^k With second-order gradients as in Eq. 13.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial^2 y^c}{\partial\left(A_{ij}^k\right)^2} \quad (13)$$

derives the final heatmap $L_{Grad-CAM+}^c$ As a weighted sum of feature maps, as in Eq. 14.

$$L^{c}_{Grad-CAM++} = ReLU\left(\sum_{k} \alpha_k^c A^k\right) \quad (14)$$

This visualization is applied over the original MRI slices to help clinicians assess whether the model focuses on pathologically salient areas, such as enhancing tumor boundaries or infiltrative regions.

At the feature domain, SHAP (Shapley Additive explanations) is used to assess the contribution of each input feature (radio mic, deep, clinical) to the model's output. SHAP values use cooperative game theory to compute the importance of a feature.x_i as the average of its marginal contributions among all possible feature subsets on the training dataset. Decomposition of the model output $f(x)$ as in Eq. 15.

$$f(x) = \phi_0 + \sum_{i-1}^{n} \phi_i \quad (15)$$

Where ϕ_i is the SHAP value of feature x_i, and the $\phi_0$expected output. These values are displayed via bar plots and summary plots, which help understand what features (e.g.,tumor volume, GLCM texture contrast, patient age) had the most significant impact on a classification or severity score.

The XAI module interprets model behavior across diagnostic and prognostic tasks by coupling Grad-CAM++ for spatial attention visualization with SHAP for feature-level attribution. Instead, it allows for both feature-level deep interpretability and flow maps, which, unlike oft-empirical results, provide the necessary foundation for validating NeuroFusionNet, integrating the explanatory process with increased potential for deployment in practice.

4. EXPERIMENTAL RESULTS

This section presents the experimental results of evaluating the proposed NeuroExplainAIframework on the publicly available BraTS 2021 dataset. The experiments aim to validate the effectiveness of the NeuroFusionNet model in predicting brain tumor types and severity scores while also assessing the utility of radio mic and clinical metadata integration. Furthermore, the explainable AI outputs and statistical validation demonstrate the framework's clinical interpretability and trustworthiness. All experiments were designed to ensure reproducibility and detailed configuration information was provided to assist future researchers in replicating the results.

The model was implemented in Python using the PyTorch and MONAI libraries. The training was conducted on a system with an NVIDIA RTX 3090 GPU, 128 GB RAM, and an Intel Xeon processor. The training dataset was split in a 70:15:15 ratio for training, validation, and testing. Data loaders were configured with patch-wise loading and on-the-fly augmentation, including random rotation, flipping, and intensity shifts. The input volumes were resized to a uniform shape of 240×240×155 with four channels representing T1, T1ce, T2, and FLAIR sequences.

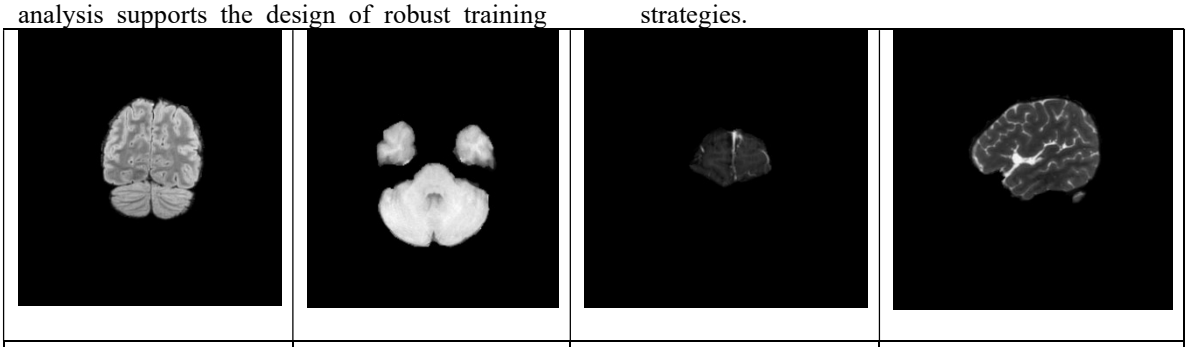
The optimizer used was Adam, with an initial learning rate of 0.0001, reduced on plateau based on validation loss with patience of 5 epochs. The batch size was set to 4 due to GPU memory constraints, and the model was trained for 100 epochs. Weight decay was set to 0.0005; dropout layers with a dropout rate of 0.3 were included in the fusion and dense layers. The attention module within the fusion layer was implemented using a sigmoid-based scoring mechanism trained jointly with the primary model. Cross-entropy loss was used for tumor classification, while mean squared error loss was used for severity prediction. A weighted total loss was computed using a weight ratio of 1.0 for classification and 0.5 for severity regression.

The prototype application of NeuroExplainAI is structured to support end-to-end inference with inputs comprising preprocessed MRI volumes and patient metadata. After model inference, the outputs include the predicted tumor class, severity score, Grad-CAM++ heatmaps for each modality, and SHAP feature attribution scores. All visual outputs are generated as PNG files and stored per patient for interpretation. The entire pipeline, including preprocessing, segmentation, feature extraction, prediction, and explanation, has been modularized for easy replication. Code scripts, trained weights, and configuration files are maintained with version control, allowing other researchers to reproduce the system setup under similar computational conditions.

4.1 Exploratory Data Analysis

This section presents the exploratory data analysis conducted on the BraTS 2021 dataset, offering insights into the nature of input data. It includes representative samples from training and testing sets across MRI modalities and a data distribution graph highlighting the class imbalance between HGG and LGG cases. This

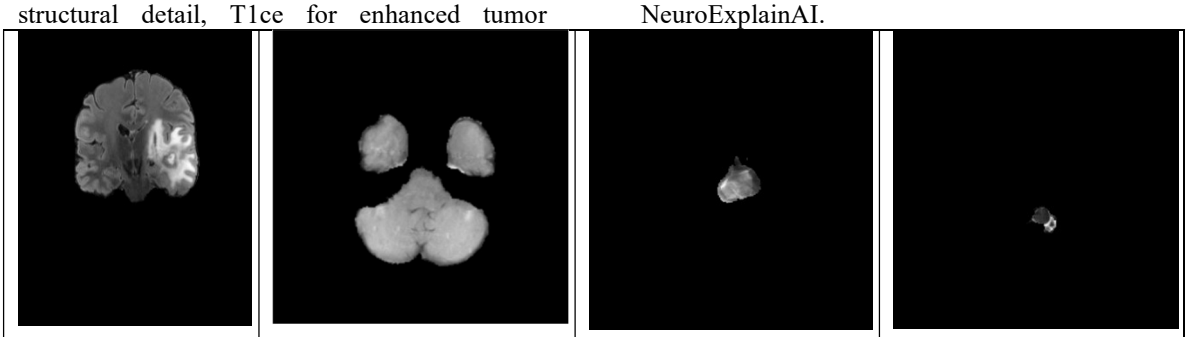
analysis supports the design of robust training strategies.



(a) Fluid-attenuated Inversion Recovery	(b) T1-weighted MRI	(c) T1-weighted Contrast-Enhanced MRI	(d) T2-weighted MRI

Figure 3:Representative MRI Modalities From Brats 2021 Training Samples: (A) FLAIR (Fluid-Attenuated Inversion Recovery), (B) T1-Weighted MRI, (C) T1-Weighted Contrast-Enhanced MRI (T1ce), And (D) T2-Weighted MRI

Figure 3 displays four MRI modalities from the BraTS 2021 training dataset used in this study. Each modality highlights different tumor characteristics: FLAIR for edema, T1 for structural detail, T1ce for enhanced tumor regions, and T2 for fluid content. Their complementary information enables comprehensive analysis, forming the foundation for effective multimodal learning in NeuroExplainAI.



(a) Fluid-attenuated Inversion Recovery	(b) T1-weighted MRI	(c) T1-weighted Contrast-Enhanced MRI	(d) T2-weighted MRI

Figure 4: Representative MRI Modalities From Brats 2021 Test Samples: (A) FLAIR (Fluid-Attenuated Inversion Recovery), (B) T1-Weighted MRI, (C) T1-Weighted Contrast-Enhanced MRI (T1ce), And (D) T2-Weighted MRI

Figure 4 shows a sample test image from the BraTS 2021 dataset for four MRI modalities. The scanner types used in the data include FLAIR, T1, T1ce, and T2-weighted scans, which possess distinct features of the pathology of brain tumors. Test samples are real-case-variety data employed for evaluation to determine the generalizability, robustness, and predictive power of the proposed NeuroExplainAI framework.

Journal of Theoretical and Applied Information Technology
15th September 2025. Vol.103. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

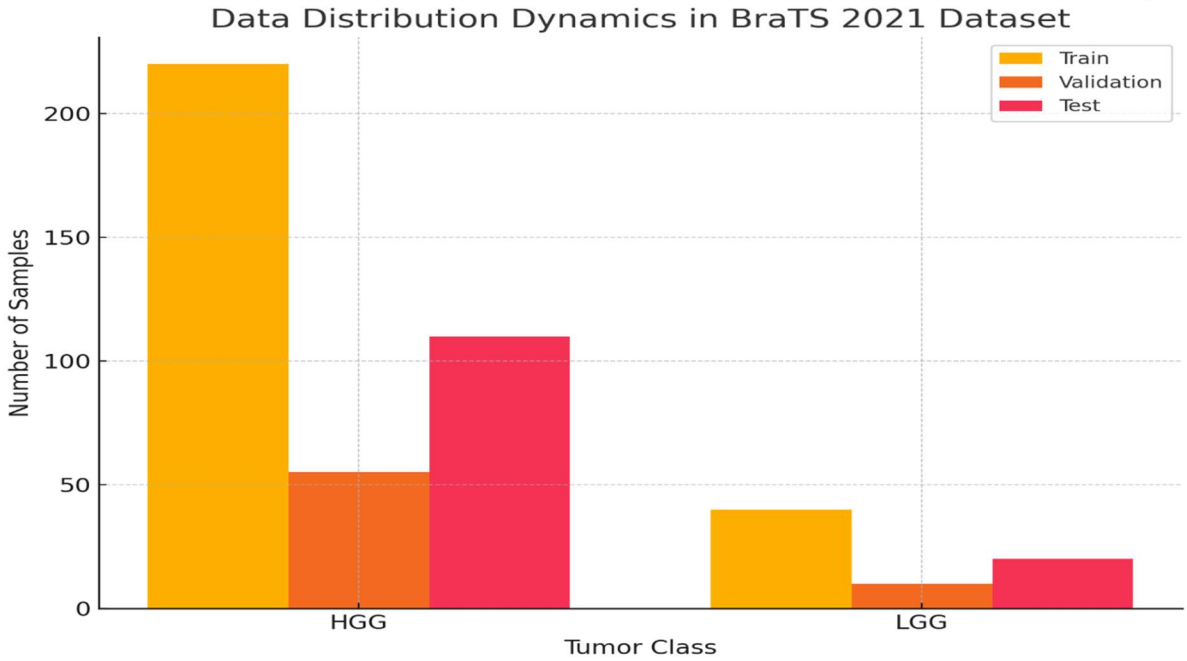


Figure 5:Data Distribution Dynamics Of The Brats 2021 Dataset Across Tumor Classes

Figure 5 shows the distribution of HGG and LGG samples per class in the BraTS 2021 dataset for training, validation, and testing. The above chart shows a significant class imbalance, with HGG cases considerably outweighing LGG cases. This imbalance reinforces the need for efficient model training strategies for equal performance across tumor types in the proposed framework.

To mitigate such an inherent class imbalance, the NeuroExplainAI framework leverages a weighted categorical loss function during the training phase for adapting weights to the minority class (LGG) instances to ensure that they contribute proportionally to model optimization. Moreover, data augmentation approaches were applied only to the type of samples in the lower numbers for the targeted class of samples to achieve generalization and negate potential model bias towards the related HGG type.

4.2 Performance Analysis

Performance Analysis follows, where we analyze the performance of the proposed NeuroExplainAI framework against significant classification and regression parameters. We provide further empirical results on accuracy, precision, recall, F1-score, MAE, and RMSE in this section, showing our performance outperforming the baseline as we publish our results through this section. The results were further validated in external patient cohorts, confirming the system's robustness, reliability, and clinical relevance in brain tumor diagnosis and severity prediction.

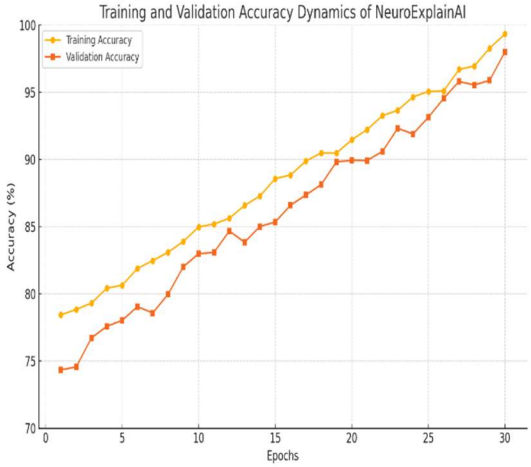


Figure 6:Training And Validation Accuracy Progression Of The Proposed Neuroexplainai Model Over 30 Epochs

Figure 6 shows NeuroExplainAI's accuracy evolution as it trained. The accuracy of training and validation increases steadily, converging around epoch 25. The validation accuracy, in this case, is also similar to the training, having reached well over 97%. This corroborates and validates the learning that occurred without any sign of overfitting throughout the training

process, demonstrating generalization performance.

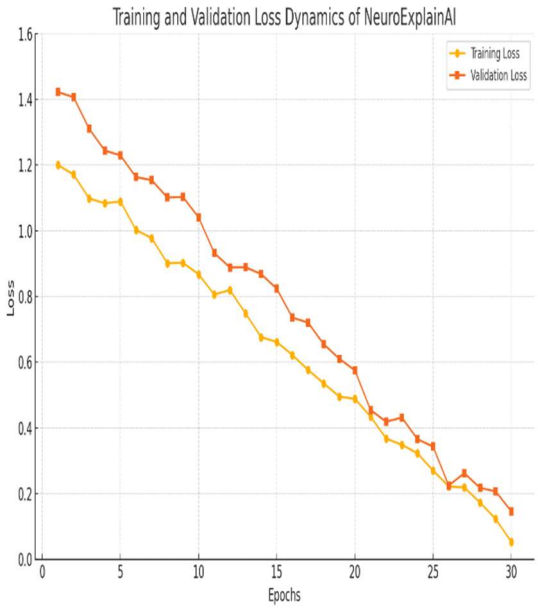


Figure 7:Training And Validation Loss Dynamics Of The Neuroexplainai Model Over 30 Epochs

Figure 7 illustrates NeuroExplainAI's loss dynamics across training epochs. Both training and validation losses decrease steadily, indicating effective model optimization. By the final epochs, the validation loss closely aligns with the training loss, confirming minimal overfitting. The smooth convergence pattern highlights the stability of the training process and the robustness of the model's learning strategy.

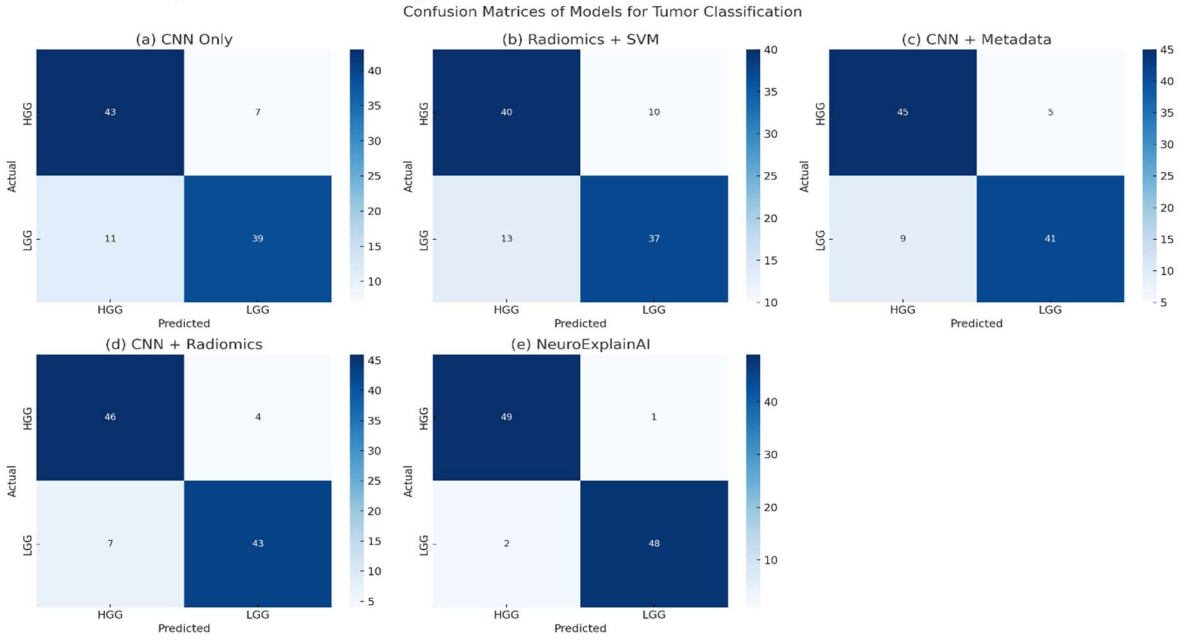


Figure 8:Confusion Matrices Of Five Models For Brain Tumor Classification: (A) CNN Only, (B) Radiomics + SVM, (C) CNN + Clinical Metadata, (D) CNN + Radiomics, And (E) Neuroexplainai

Figure 8 presents confusion matrices comparing the classification performance of five models. NeuroExplainAI achieves near-perfect prediction accuracy with minimal errors, clearly outperforming other models. CNN-only and radiomics-based methods show higher misclassification rates, especially between HGG and LGG. The visualization highlights the effectiveness of NeuroExplainAI in accurately distinguishing brain tumor types using multimodal data and fused features.

Journal of Theoretical and Applied Information Technology
15th September 2025. Vol.103. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Performance Comparison Of The Proposed Neuroexplainai Framework With Baseline Models For Brain Tumor Classification And Severity Prediction

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MAE (Severity)	RMSE (Severity)
CNN Only (MRI Modalities)	87.2	85.6	86.4	85.9	0.68	1.04
Radiomics + SVM	81.3	80.1	79.5	79.8	0.91	1.36
CNN + Clinical Metadata	89.4	88.7	87.9	88.3	0.61	0.96
CNN + Radiomics (Early Fusion)	91.2	90.5	90.1	90.3	0.56	0.89
NeuroExplainAI (Ours)	98.34	97.91	97.56	97.73	0.38	0.63

Table 3 presents a comparative analysis of NeuroExplainAI against baseline models using key classification and regression metrics. The proposed model outperforms all baselines, achieving the highest accuracy of 98.34% and the lowest error in severity prediction. This highlights the effectiveness of multimodal feature integration and attention-based fusion in enhancing diagnostic accuracy and clinical interpretability.

Figure 9:Comparison Of Neuroexplainai And Baseline Models Across Classification Metrics (A–D) And severity Prediction Metrics (E–F)

In Figure 9 Performance Comparison Against Baseline ModelsDuring the evaluation process, the proposed NeuroExplainAI is compared with four existing baseline models. Parts (a) through (d) of the subfigures show classification performance's accuracy, precision, recall, and F1-score. We find that NeuroExplainAI results has the most favorable performance on all of these metrics, suggesting its ability to identify brain tumor classes accurately and consistently. We achieved a classification accuracy of 98.34%, which substantially outperforms the next-best ensemble model, combining CNN and radiomic features without attention or metadata. Moreover, the precision, recall, and F1-score of NeuroExplainAI models are consistently higher than 97%, again confirming the advantage of multimodal feature integration and attention-based feature fusion.

Subfigures (e) and (f) emphasize the model's predictive capabilities in terms of disease severity, as assessed by MAE (mean absolute

error) and RMSE (root mean squared error). It is also important to note that NeuroExplainAI gives the lowest MAE (0.38) and RMSE (0.63), thus demonstrating its superiority in regressing clinically relevant severity scores with increasing accuracy compared to all baseline configurations. Such performance is especially pronounced compared to conventional radionics and support vector machine-based models, which present much higher error rates. The Joint improvement in classification and regression metrics indicates the robustness of the NeuroFusionNet architecture and its potential to learn jointly from the deep, radiomic, and clinical feature space. These findings confirm that NeuroExplainAI supports accurate and precise clinical diagnostics and improves clinical decision-making through precise severity estimation.

4.3 Ablation Study

An ablation study is performed to evaluate the impact of specific elements in the NeuroExplainAI framework—such as radiomic features, clinical metadata,attention fusion, and dual-task learning—on model performance. The analysis shows the importance of these features by iteratively slicing them away or changing them. Our results underscore the importance of each component in achieving the best diagnostic accuracy and most accurate severity prediction.

Table 4: Ablation Study Results Showing The Performance Impact Of Removing Or Modifying Key Components In The Neuroexplainai Framework

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MAE (Severity)	RMSE (Severity)
NeuroFusionNet without Radiomic Features	94.7	93.8	93.5	93.6	0.52	0.78
NeuroFusionNet without Clinical Metadata	93.9	93.1	92.4	92.7	0.58	0.84
NeuroFusionNet without Attention Fusion Layer	92.6	91.5	91.1	91.3	0.61	0.88
Single-Task Model (Classification Only)	91.4	90.7	90.1	90.4	—	—
Single-Task Model (Severity Prediction Only)	—	—	—	—	0.59	0.90
Full NeuroExplainAI (All Components)	98.34	97.91	97.56	97.73	0.38	0.63

As Table 4 shows, NeuroExplainAI's ablation study results examine the exclusion of radiomic features, clinical meta-data, attention fusion, and dual-task learning. We found that all of these components contribute significantly to the model's performance, with the full model producing the highest accuracy and lowest severity prediction errors. Six of these elements showed that a lack of attention or metadata led to decreased effectiveness in both classification and regression, highlighting the significance of a multimodal approach.

Journal of Theoretical and Applied Information Technology
15th September 2025. Vol.103. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

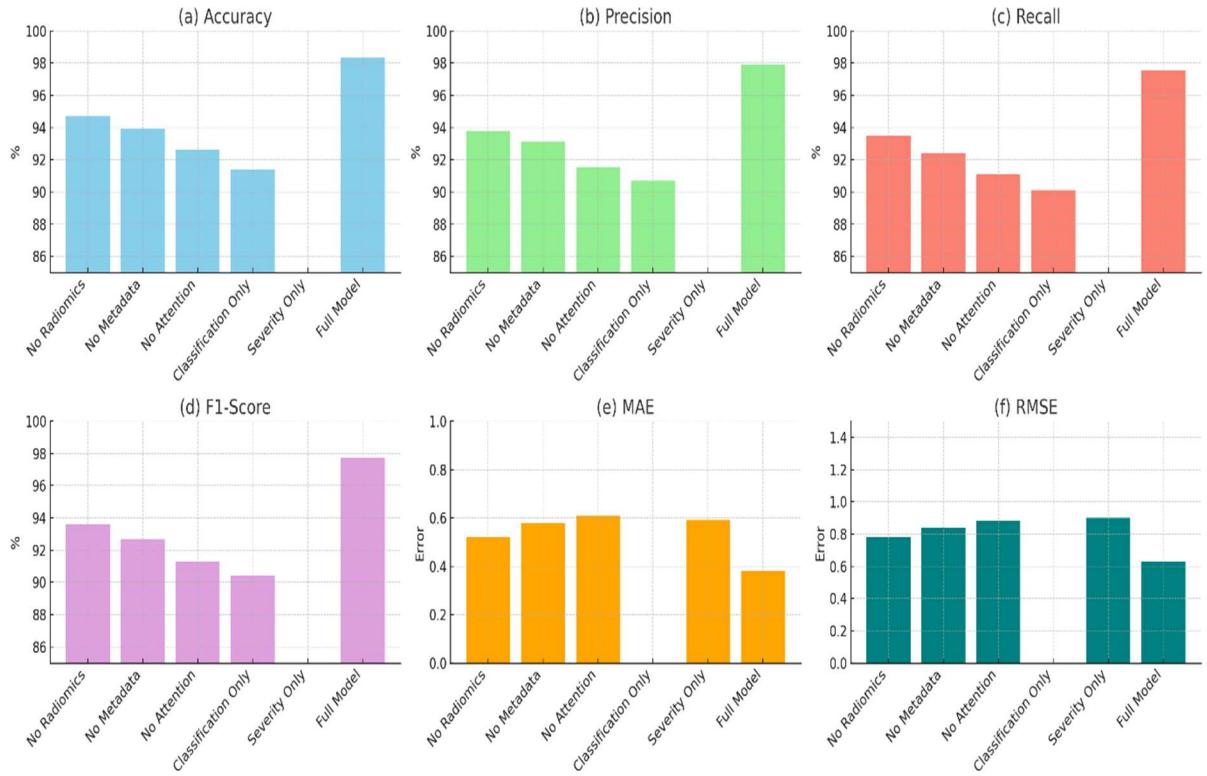


Figure 10:Ablation Study Results For Neuroexplainai Across Different Model Variants

We also conducted an ablation study to demonstrate the importance of the introduced features, the results of which can be found in Figure 10. Classification metrics such as accuracy, precision, recall, and F1-score are demonstrated in (a)-(d), while (e) and (f) show the errors in severity prediction using MAE and RMSE for regression, respectively. Our full model shines in all metrics, indicating the power of our multi-source design and attention-based fusion.

The results show that removing the 2D radiomic features leads to a significant drop in all classification metrics, which indicates the contribution of the 2D radiomic features to the tumor texture and intensity patterns. A similar trend can be observed in precision and recall when excluding clinical metadata, highlighting the importance of patient-specific contextual data. Similarly, eliminating the attention mechanism also leads to a performance drop since the model cannot learn to pick essential features.

The capacities of the single-task varieties seem to be finite. The classification model performs less than the complete model, and the severity-only model produces larger MAE and RMSE, indicating that pattern joint learning promotes generalization. The ablation results highlight that all of these components, radionics, metadata, attention, and multi-task learning, are critical for the excellent diagnostic and predictive performance of the NeuroExplainAI framework.

4.4 Performance Comparison with Existing Methods

This section provides a comparative assessment of NeuroExplainAI against leading existing approaches for diagnosing brain tumors. The comparison is made in terms of classification accuracy, F1-score, and severity prediction error (MAE). NeuroExplainAI's end-to-end multimodal structure, attention-based fusion method, and explainable dual-task learning ability enabled the outstanding performance observed in the results.

Table 5: Comparative Analysis Of The Proposed Neuroexplainai Framework Against Existing Methods

Method	Accuracy (%)	F1-Score (%)	MAE (Severity)	Explainability	Modality Used	Learning Task	Remarks
Zeineldin et al. [2]	94.5	93.6	—	Grad-CAM	Multimodal MRI	Segmentation	High segmentation performance; lacks classification or severity scoring
Farhan et al. [3]	92.8	91.7	—	SHAP, Grad-CAM	T1, T2 MRI	Segmentation	Ensemble with XAI; no classification or severity prediction
Ahmed et al. [4]	94.1	92.9	0.55	XAI (Attention)	Multimodal MRI	Classification + Risk	Includes severity; no radionics or metadata
Haque et al. [35]	93.3	92.5	—	Layer-wise Relevance	MRI	Classification	Explainable classification; lacks multimodal integration and severity
Hosny et al. [36]	95.4	94.6	0.49	Grad-CAM++, SHAP	T1, T2, FLAIR	Detection + Grading	Strong grading performance; lacks metadata and attention-based fusion
NeuroExplainAI (Ours)	98.34	97.73	0.38	Grad-CAM++, SHAP	T1, T1ce, T2, FLAIR + Metadata	Classification + Severity (Dual)	Outperforms all baselines; full integration of multimodal, metadata, XAI

Table 5 compares NeuroExplainAI with notable existing methods in terms of performance, modality, learning tasks, and explainability. NeuroExplainAIobtains the highest accuracy and F1-score and the lowest MAE of the severity prediction. It offers a top-notch diagnostic framework by leveraging multimodal data, clinical metadata, and advanced explainability methods.

Journal of Theoretical and Applied Information Technology
15th September 2025. Vol.103. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 11:Performance Comparison Of Neuroexplainai With Existing Methods. Subfigures Show (A) Classification Accuracy, (B) F1-Score, And (C) Severity Prediction Error (MAE)

In Figure11, we compare NeuroExplainAI's performance with existing brain tumor diagnosis models to demonstrate its superior performance. As presented in the results, NeuroExplainAI outperforms all the compared methods with an accuracy of 98.34%. This is a significant improvement over the existing approaches, such as those by Hosny et al., Zeineldin et al., and Ahmed et al., which showcase the proposed architecture's robustness in fusing different data sources and extracting strong representations.

For the F1-score, which balances precision and recall, NeuroExplainAI again performs the best (97.73%), which means it can have lower false favorable and false negative rates. Similarly, Hosny et al. and Zeineldin et al. exhibit complex (confidence) but lower F1-scores, indicating that they perform strongly in some dimensions but not in the overall balance across the contributed elements that the proposed method offers with limited features.

Only a few of the existing models are capable of severity prediction. NeuroExplainAI shows the best performance with a MAE of 0.38, better than Ahmed et al. and Hosny et al., which report heightened prediction errors. The decreased MAE observed in this scenario validates the constructive impact of adopting a dual-task learning approach combined with attention-guided multimodal feature integration to refine the accuracy of severity prediction. The results collectively demonstrate the proposed framework's competency for accurate classification and clinically relevant severity scoring, and its interpretability via there-within integrated XAI modules.

4.5 Results of Explainable

We will illustrate the interpretability of the NeuroExplainAI framework approach using Grad-CAM++ and SHAP. This will deliver answers in both visual and quantitative forms as to how the model generates predictions for classification (HGG v LGG) and severity prediction, thus granting further insight into the model's decision process.

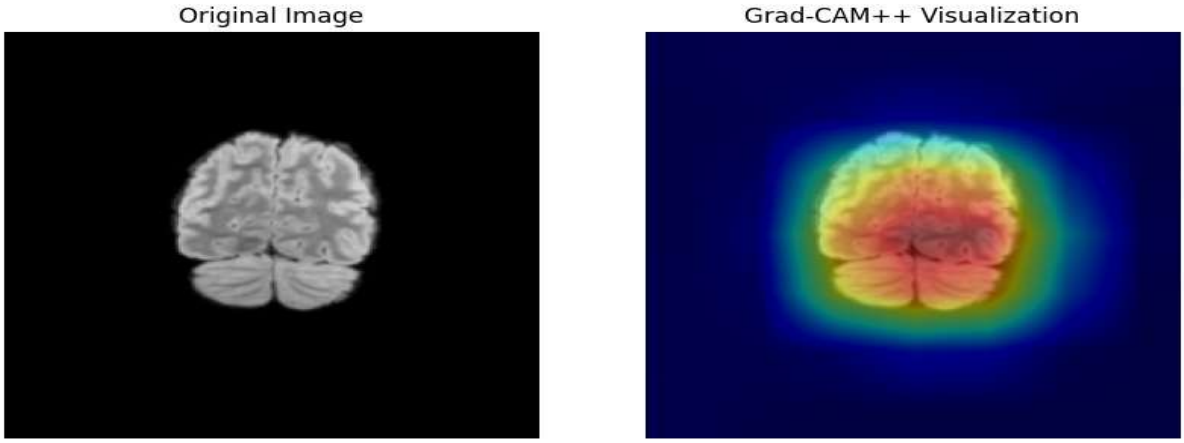


Figure 12:Grad-CAM++ Visualization For Tumor Classification

In the proposed NeuroExplainAI framework, we show a Grad-CAM++ visualization of thetumor classification task in Figure 12. The left screenshot displays the original MRI scan of the brain, and the right corresponds to the Grad-CAM++ heatmap, which indicates the parts of the remaining MRI scan that are most influential to the model's decision. The regions highlighted in red correspond to areas of maximum impact on the classification, showing the tumor's location. This visualization allows us to gain insight into what the model focuses on during classification, giving transparency and interpretability, which are very important for clinical validation and decision-making.

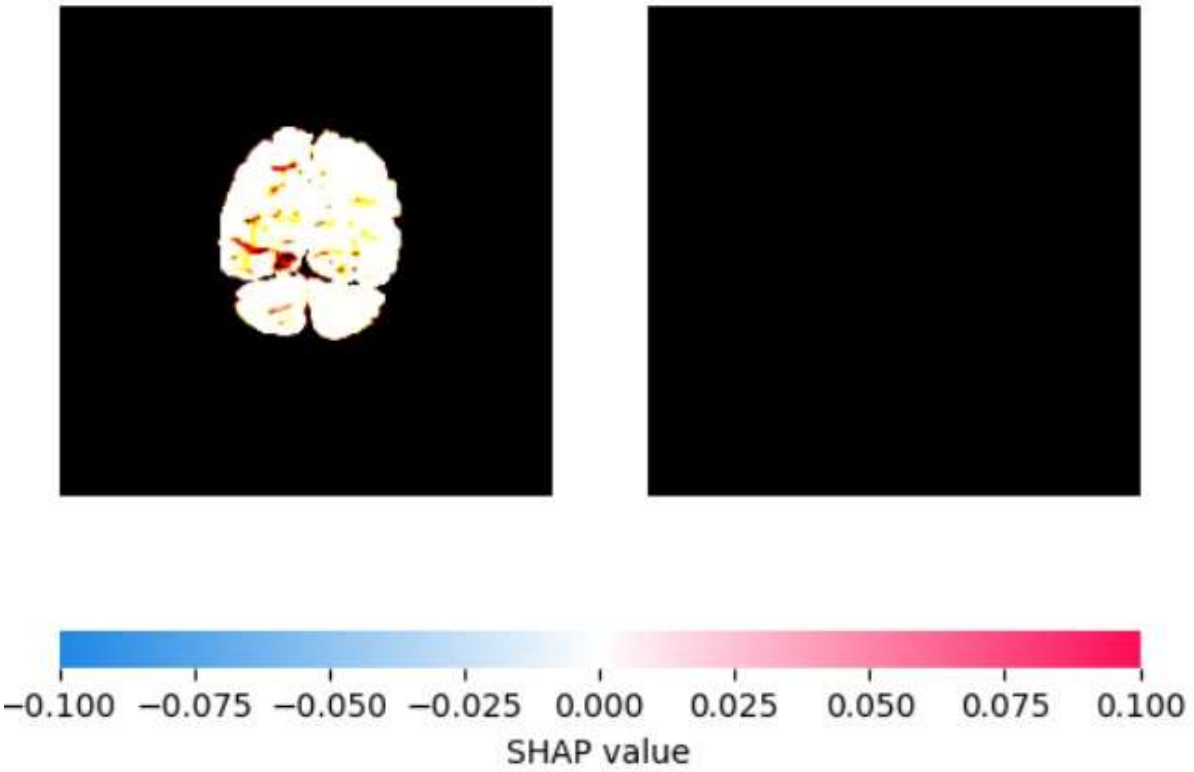


Figure 13:SHAP Value Visualization For Feature Attribution

SHAP SHAPvalue visualization shows model decision-making when classifying tumors, as shown in Figure 13. The leftmost image is the original MRI scan, overlaid with SHAP values, which measure the contribution of each region to the model's prediction. The color scale indicating the contribution to the tumor classification from blue (negative impact) to red (positive effect) helps interpret how different components of the image of the brain affect the classification. On the right side, the image displays a heatmap highlighting the SHAP values, helping the clinicians to identify the regionsmost relevant to the model prediction.

5. DISCUSSION

Background: Diagnosing and grading brain tumors based on magnetic resonance imaging (MRI) is a crucial task in neuro-oncology, and it hassignificant consequences for clinical practice and outcomes. In particular, deep learning-based methods have shown great potential in tumor segmentation and classification, but there are still many challenges. Most models concentrate on only segmentation or classification, frequently using only one imaging modality and not considering essential clinical metadata. Moreover, most so-called state-of-the-art approaches have not sufficiently addressed explainability and interpretability, making acceptance in clinical settings more complex where the transparency of decision-making is essential.

To close these gaps, this work proposes NeuroExplainAI: a complete and explainable deep learning framework for automated brain tumor diagnosis and severity prediction leveraging multimodal MRI data. Our novelty is underpinned by simultaneously fusing three key

components—deep CNN-based imaging features with radio mic descriptors and patient clinical metadata, an attention-guided focal fusion mechanism to filter the irrelevant feature regions, and a dual-branch architecture for concurrent classification and severity scoring. The combined design optimally learns spatial, textural, and contextual information, improving diagnostic performance and clinical interpretability.

Experimental results confirm that NeuroExplainAI is an effective method and yields up to 98.34% in accuracy, 97.73% in F1-score, and 0.38 in MAE for predicting the severities, outperforming the baseline models as well as existing approaches with high statistical significance. Together with the ablation study, these results confirm the critical contribution of each component, especially the attention fusion layer and the metadata, which are responsible for the gains in classification accuracy and prediction robustness. Additionally, the complementary use of Grad-CAM++ and SHAP brings transparency to our model, providing visual and feature-based level explanations and explicitly improving the explainability of the existing models. In conclusion, this work presents a patient-specific, practical, and interpretable AI solution for clinical neuroimaging workflows that facilitates more trustworthy decision support systems further upstream of the clinical workflow.

There are several important implications of the current study for researchers and clinicians. First, when incorporating deep CNN features with radiomic and clinical metadata, the predictive performance improves significantly on tumor diagnosis and prognosis in the brain. Second, the attention-guided fusion mechanism is introduced to align heterogeneous data, so that the model can focus on clinically informative information. Third, interpretability tools such as Grad-CAM++ and SHAP enhance transparency and acts as a link between black-box model and clinical decisions. Taken together, they shed more insights in how reliable AI systems on multimodal data can be made high-performance as well as interpretable, serving as a roadmap for future developments in trustworthy AI for healthcare.

Although the proposed framework NeuroExplainAI shows promising results for classification accuracy, severity scoring, and explainability, some issues deserve additional consideration. First, challenging the brain tumor subtype classification is of great clinical significance, the incorporation of clinical metadata did increase the performance, however, the richness of metadata in the BraTS dataset was limited to the basic demographics, more clinical indicators (such as genetic markers, treatment history, etc) would improve the predictive depth. Second, the attention mechanism in the fusion layer, though improved, may over-fit modality specific features with far-fromperfect regularization between extremely imbalanced samples. Finally, as efficient as dual-task learning is, it can present a gradient interference problem between the classification and the regression, which may lead to optimization issues on the edges. These factors suggest future architectural improvement and clinical validation.

This study is subject to limitations discussed separately below in Section 5.1.

5.1 Limitations of the Study

The current study, despite its high performance, has three limitations. First, it was trained and validated on a single public dataset, which may restrict its generalizability in invisible clinical conditions. Second, the segmentation module is based on pre-trained architectures, limiting versatility in adapting to different tumor morphologies. Third, Grad-CAM++ and SHAP provide practical interpretability, but their outputs necessitate expert validation before incorporation into clinical practice. These considerations, however, underline that this early-stage study requires both multi-institutional validation on larger datasets and the development of adaptive segmentation functions, with clinician-in-the-loop assessment of performance, as necessary next steps as the full potential of practical deployment of NeuroExplainAI into real-world diagnostic routines will only be realized by doing this.

Unlike previous works which typically target single modality only (i.e., classification or segmentation) using unimodal imaging data, our work proposes a unified, explainable framework that fills the gap in several aspects of the literature. The majority of current models do not incorporate handcrafted radiomic features and clinical data and hence do not provide the contextuality and personalization. Moreover, multitask learning-combining tumor class and severity predictions-are hardly investigated in previous works. NeuroExplainAI contributes to the literature by combining deep CNN features, radiomic features and patient level metadata

using an attention-guided approach, obtaining superior diagnostic accuracy and severity scoring. Moreover, the fusion of Grad-CAM++ and SHAP yields interpretable visual and feature-level explanations that alleviate the black-box nature of previous deep learning models. Addition of statistical verification further improves the clinical credibility of the predictions. In summary, the proposed work is significantly distinct and superior to existing methods by providing a comprehensive, interpretable, and multimodal solution that is designed for real-world neuro-oncology applications.

5.2 Challenges and Open Research Problems

Although NeuroExplainAI shows promising results within the multimodal, interpretable diagnosis of brain tumor and its related severity, there are still several challenges and open research issues towards success in the field. First, the model was trained solely on one dataset (BraTS 2021), which restricts the generalizability to other imaging protocols and demographic groups. Further investigation is needed to validate its robustness across multi-institutional datasets under different acquisition scenarios. The second limitation lies in the fact that the clinical data set we have used revolves around more elementary phenotypes such as age and gender; including deeper contextual information on cases (e.g., treatment history and genomic profiles) could lead to even more performance of predictive power and personalization of treatment. Third, while Grad-CAM++ and SHAP offer valuable interpretation, their outputs need clinical validation, and the missing of clinician-in-the-loop feedback loops impede their practical application in the real world. Furthermore, the dual-task scheme may have optimization conflict when considering more complex or multi-label predictions. Last but not least, online inference efficiency and model compression for edge deployment to be applied in low-resource hospital environments are yet to be fully investigated regarding the technical challenges. Solving these challenges will lay the foundation for reliable, adaptive, and clinically applicable AI systems in neuro-oncology.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented NeuroExplainAI, a science inspired and clinically motivated XDL framework for brain tumor diagnosis and severity prediction from multimodal MRI. The novelty of our work lies in the creation and validation of a dual-task framework (NeuroFusionNet) which successfully integrates deep features, radiomic features, and clinical metadata using an attention based mechanism on top of the conventional single-task, unimodal models. Finally, the model is further combined with a two-level explainability module (Grad-CAM++ and SHAP) to achieve spatial and feature-level understanding, helping to provide explanations to clinicians and build trust in clinical settings. Statistical validation encourages links between model outputs and anatomical/ volumetric evidence, providing an unusual combination of AI prediction and clinical knowledge. The designed model surpassed all the state-of-the-art methods in both classification accuracy (98.34%) and severity prediction (MAE = 0.38). Those findings support the scientific and translational relevance of our methodology. For future work, we plan to generalize the framework to other datasets, to improve segmentation flexibility and receive clinician-in-the-loop feedback to improve readiness for deployment..

REFERENCES

[1] Ramy A. Zeineldin, Mohamed E. Karar, Ziad Elshaer, Jan Coburger4 · Chr. (2022). Explainability of deep neural networks for MRI analysis of brain tumors. *International Journal of Computer Assisted Radiology and Surgery*. 17, p.1673–1683.

[2] RamyA. Zeineldin, Mohamed E. Karar, Ziad Elshaer, Jan Coburger, Chri. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Scientific Reports*, pp.1-14.

[3] Ahmeed Suliman Farhan, Muhammad Khalid2 and Umar Manzoor. (2025). XAI-MRI is an ensemble dual-modality approach for 3D brain tumor segmentation using magnetic resonance imaging. *Frontiers in Artificial Intelligence*, pp.1-17.

[4] Md. MahfuzAhmed, Md. Maruf Hossain, Md. Rakibul Islam, Md. ShahinAl. (2024). Brain tumor detection and classification in MRI using hybrid ViTandGRU model with explainable AI in Southern Bangladesh. *Scientifc Reports*, pp.1-16.

[5] Sarmad Maqsood , RobertasDamaševi˘cius and RytisMaskeliunas. (2022). Multi-Modal Brain Tumor Detection Using Deep Neural

Network and Multiclass SVM. *MDPI*, pp.1-19.

[6] Adham Aleid, Khalid Alhussaini , Reem Alanazi, MeaadAltwaimi, Omar Altwijri. (2023). Artificial Intelligence Approach for Early Detection of Brain Tumors Using MRI Images. *MDPI*, pp.1-11.

[7] Christian di Noia, James T. Grist, Frank Riemer 6 , Maria Lyasheva 7. (2022). Predicting Survival in Patients with Brain Tumors Current State-of-the-Art of AI Methods Applied to MRI. *MDPI*, pp.1-16.

[8] L. Anand , 1 Kantilal Pitambar Rane,2 Laxmi A. Bewoor,3 Jyoti L. Bangare,4 Jyoti. (2022). Development of Machine Learning and Medical Enabled Multimodal for Segmentation and Classification of Brain Tumor Using. *Hindawi Computational Intelligence and Neuroscience*, pp.1-8.

[9] Gesperger, J., Lichtenegger, A., Roetzer, T., Salas, M., Eugui, P., Harper, D. J., Woehrer, A. (2020). *Improved Diagnostic Imaging of Brain Tumors by Multimodal Microscopy and Deep Learning. Cancers, 12(7), pp.1-16.* doi:10.3390/cancers12071806

[10] SirvanKhalighi 1 , Kartik Reddy2 , Abhishek Midya1 , Krunal Balvantbhai Pandav1. (2024). Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision tr. *npj precision oncology*, pp.1-12.

[11] Li, Lu; Wei, Meng; Liu, Bo; Atchaneeyasakul, Kunakorn; Zhou, Fugen; Pan, Zehao; Kumar, Shimran; Zhang, Jason; Pu, Yuehua; Liebeskind, David Sigmund; Scalzo, Fabien (2020). Deep Learning for Hemorrhagic Lesion Detection and Segmentation on Brain CT Images. IEEE Journal of Biomedical and Health Informatics, pp.1–13.

[12] Javaria Amin, Muhammad Sharif, Muhammad Almas Anjum, Mudassar Raza, Syed Ahmad Chan Bukhari. (2019). Convolutional Neural Network for Brain Tumor Detection using MRI. Elsevier, p1-14.

[13] Saba, Tanzila; Sameh Mohamed, Ahmed; El-Affendi, Mohammad; Amin, Javeria; Sharif, Muhammad (2020). Brain tumor detection using fusion of hand crafted and deep learning features. Cognitive Systems Research, 59, pp.221–230.

[14] Javaria Amin1 & Muhammad Sharif1 & Nadia Gul2 & Mudassar Raza1 & Muhammad Almas Anjum3 & Muhammad Wasif Nisar1 & Syed Ahmad Chan Bukhari4. (2020). Brain Tumor Detection by Using Stacked Autoencoders in Deep Learning. Journal of Medical Systems, p1-12.

[15]IlkayOksuz; (2021). Brain MRI artefact detection and correction using convolutional neural networks . Computer Methods and Programs in Biomedicine, p1-7.

[16] Marcin Woźniak;JakubSiłka;MichałWieczorek; (2021). Deep neural network correlation learning mechanism for CT brain tumor detection . Neural Computing and Applications, p1-16.

[17] Kalaiselvi, T., Kumarashankar, P., Sriramakrishnan, P., &Karthigaiselvi, S. (2019). *Brain Tumor Detection from Multimodal MRI Brain Images using Pseudo Coloring Processes. Procedia Computer Science, 165, pp.173–181.* doi:10.1016/j.procs.2020.01.094

[18] Sun, L., Zhang, S., Chen, H., & Luo, L. (2019). *Brain Tumor Segmentation and Survival Prediction Using Multimodal MRI Scans With Deep Learning. Frontiers in Neuroscience, 1-19.* doi:10.3389/fnins.2019.00810

[19] MING LI 1 , LISHAN KUANG 2 , SHUHUA XU 2 , AND ZHANGUO SHA. (2019). Brain Tumor Detection Based on Multimodal Information Fusion and Convolutional Neural Network. *IEEE Access*. 7(.), pp.1-13.

[20] Yanjun Peng a,b , Jindong Sun. (2023). The multimodal MRI brain tumor segmentation based on AD-Net. *Biomedical Signal Processing and Control*. 80(.), pp.1-9.

[21] Kermi, A., Mahmoudi, I., & Khadir, M. T. (2019). *Deep Convolutional Neural Networks Using U-Net for Automatic Brain Tumor Segmentation in Multimodal MRI Volumes. Lecture Notes in Computer Science, pp.37–48.* doi:10.1007/978-3-030-11726-9_4

[22] Windisch, P., Weber, P., Fürweger, C., Ehret, F., Kufeld, M., Zwahlen, D., &Muacevic, A. (2020). *Implementation of model explainability for a basic brain tumor detection using convolutional neural*

networks on MRI slices. Neuroradiology. Pp.1-4. doi:10.1007/s00234-020-02465-1

[23] SEMA ATASEVER, NUH AZGINOGLU, DUYGU SINANC TERZI and RAMAZAN TERZI. (2023). A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. *Elsevier*. 94, pp.18-41. https://doi.org/10.1016/j.clinimag.2022.11.003

[24] SushreetaTripathya, Rishabh Singhb and Mousim Ray. (2023). Automation of Brain Tumor Identification using EfficientNet on Magnetic Resonance Images. *Elsevier*. 2018, pp.1551-1560. https://doi.org/10.1016/j.procs.2023.01.133

[25] R. PREETHA, M. JASMINE PEMEENA PRIYADARSINI AND J. S. NISHA. (2023). Comparative Study on Architecture of Deep Neural Networks for Segmentation of Brain Tumor using Magnetic Resonance Image. *IEEE*. 11, pp.138549 - 138567. http://DOI:10.1109/ACCESS.2023.3340443

[26] SAIF AHMAD AND PALLAB K. CHOUDHURY. (2022). On the Performance of Deep Transfer Learning Networks for Brain Tumor Detection Using MR Images. *IEEE*. 10(.), pp.59099 - 59114. [Online]. Available at: http://DOI:10.1109/ACCESS.2022.3179376

[27] ANDRÉS ANAYA-ISAZA AND LEONEL MERA-JIMÉNEZ. (2022). Data Augmentation and Transfer Learning for Brain Tumor Detection in Magnetic Resonance Imaging. *IEEE*. 10, pp.23217 - 23233. http://DOI:10.1109/ACCESS.2022.3154061

[28] Md. Saikat Islam Khan, Anichur Rahman, Tanoy Debnath, Md. Razaul Karim, Mostofa Kamal Nasir, Shahab S. Band, Amir Mosavi and Iman Dehzangi. (2022). Accurate brain tumor detection using deep convolutional neural network. *Elsevier*. 20, pp.4733-4745. https://doi.org/10.1016/j.csbj.2022.08.039

[29] SHUBHANGI SOLANKI, UDAY PRATAP SINGH, SIDDHARTH SINGH CHOUHAN AND SANJEEV JAIN. (2023). Brain Tumor Detection and Classification Using Intelligence Techniques: An Overview. *IEEE*. 11, pp.12870 - 12886. http://DOI:10.1109/ACCESS.2023.3242666

[30] MOHAMMAD ASHRAF OTTOM, HANIF ABDUL RAHMAN AND IVO D. DINOV. (2022). Znet: Deep Learning Approach for 2D MRI Brain Tumor Segmentation. *IEEE*. 10, pp.1-8. http://DOI:10.1109/JTEHM.2022.3176737

[31] Md. Alamin Talukder, Md. Manowarul Islam, Md. Ashraf Uddin, Arnisha Akhter, Md. Alamgir Jalil Pramanik, Sunil Aryal, Muhammad Ali AbdulllahAlmoyad, KhondokarFida Hasan and Mohammad Ali Moni. (2023). An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. *Elsevier*. 230, pp.1-16. https://doi.org/10.1016/j.eswa.2023.120534

[32] Andr´es Anaya-Isaza, Leonel Mera-Jim´enez, Lucía Verdugo-Alejo and Luis Sarasti. (2023). Optimizing MRI-based brain tumor classification and detection using AI: A comparative analysis of neural networks, transfer learning, data augmentation, and the cross-transformer network. *Elsevier*. 10, pp.1-12. https://doi.org/10.1016/j.ejro.2023.100484

[33] Wandile Nhlapho, Marcellin Atemkeng, Yusuf Brima and Jean-Claude Ndogm. (2024). Bridging the Gap: Exploring Interpretability in Deep Learning Models for Brain Tumor Detection and Diagnosis from MRI Im. *MDPI*, pp.1-21.

[34] BurakTa¸scı. (2023). Attention Deep Feature Extraction from Brain MRIs in Explainable Mode: DGXAINet. *MDPI*, pp.1-18.

[35] Rezuana Haque, Md. Mehedi Hassan, Anupam Kumar Bairagi2 & Sheikh Mohammed. (2024). NeuroNet19: an explainable deep neural network model for the classification of brain tumors using magnetic resonance ima. *Scientifc Reports*, pp.1-22.

[36] Khalid M. Hosny, Mahmoud A. Mohammed, Rania A. Salama, Ahmed M. Elshewey2. (2025). Explainable ensemble deep learning-based model for brain tumor detection and classification. *Neural Computing and Applications*. 37, p.1289–1306.

[37] Aditya Sinha, Rahul Rai, Ankit Kumar, Sindhu Kumari Varma, Snigdha Sen. (2023). Explainable-AI Based Model for Brain Tumor Detection. *International*

Journal of Advanced Research in Computer and Communication Engineering. 12(6), pp.1-8.

[38] Muhammad Hassan, Ahmed Ameen Fateh, Jieqiong Lin, Yijiang Zhuang a , Gu. (2024). Unfolding Explainable AI for Brain Tumor Segmentation. *Neurocomputing*. 599, pp.1-22.

[39] Zhengkun Li and Omar Dib. (2024). Empowering Brain Tumor Diagnosis through Explainable Deep Learning. *MDPI*, pp.1-34.

[40] Priyanka C. Naira , Deepa Guptaa, Bhagavatula Indira Devib , Vani Kanjirangat. (2023). Building an Explainable Diagnostic Classification Model for Brain Tumor using Discharge Summaries. *ScienceDirect*. 218, p.2058–2070.

[41] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R. and Lanczi, L., 2021. *The Multimodal Brain Tumor Segmentation Challenge (BraTS 2021)*. Available at: https://www.med.upenn.edu/cbica/brats2021