Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Swin-TCN-XAI: A Hybrid End-to-End Framework for Explainable Multi-Class Brain Tumor MRI Classification

Umme Sara

usara@niter.edu.bd

National Institute of Textile Engineering and Research (NITER) https://orcid.org/0000-0001-5519-8681

MSTMd. Mamu

National Institute of Textile engineering Research https://orcid.org/0009-0004-2472-6946

MD IRFANUL KABIR HIRA

National Institute of Textile Engineering and Research https://orcid.org/0009-0004-6718-1632

MD SOHAG HOSSAIN

National InstItution of Textile Engineering and Research https://orcid.org/0009-0003-8398-071X

Md. Kowsar Ahmed

National institute of Textile Engineering and Research https://orcid.org/0009-0003-9652-4780

Md. Mamun Ur Rashid

Federation University https://orcid.org/0009-0003-7190-1076

Research Article

Keywords: Magnetic Resonance Imaging (MRI), Swin Transformer, Temporal Convolutional Network (TCN), Brain Tumor Classification, Explainable Artificial Intelligence (XAI), Patient-Wise Evaluation

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10644934/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Swin-TCN-XAI: A Hybrid End-to-End Framework for Explainable Multi-Class Brain Tumor MRI Classification

ARTICLE INFO

Keywords:
Magnetic Resonance Imaging (MRI)
Swin Transformer
Temporal Convolutional Network (TCN)
Brain Tumor Classification
Explainable Artificial Intelligence (XAI)
Patient-Wise Evaluation

ABSTRACT

Brain tumor classification from Magnetic Resonance Imaging (MRI) is a clinically important yet challenging task due to substantial inter-patient variability, overlapping imaging characteristics among tumor subtypes, and the limited interpretability of deep learning models in safety-critical medical applications. This study proposes Swin-TCN-XAI, a hybrid end-to-end deep learning framework, which integrates a Swin Transformer and a Temporal Convolutional Network (TCN) to jointly learn complementary spatial and sequential feature representations for multi-class brain tumor classification. The Swin Transformer captures hierarchical spatial information through shifted-window self-attention, while the TCN models inter-token dependencies using dilated causal convolutions. Both components are optimized simultaneously through end-to-end fine-tuning with differential learning rates to improve feature adaptation and classification performance. To enhance clinical transparency, the framework incorporates a multi-level explainability strategy combining Grad-CAM for lesion localization, SHAP for global and class-specific feature attribution, and LIME for patient-level decision interpretation. The proposed framework was evaluated on a four-class brain MRI dataset (glioma, meningioma, pituitary tumor, and no tumor) using a strict patient-wise train-test split to eliminate data leakage. The model achieved 99.41% accuracy, with macro-averaged precision, recall, and F1-score of 0.9940, 0.9948, and 0.9944, respectively, outperforming state-of-the-art transformer, temporal, and hybrid baselines. Class-wise analysis demonstrated consistently high performance, including perfect precision and recall for the no-tumor class and an F1-score of 0.9925 for the clinically challenging glioma class. These results demonstrate that the proposed Swin-TCN-XAI framework provides an accurate, robust, and interpretable approach for automated brain tumor diagnosis, supporting trustworthy clinical decision-making in MRI-based neuro-oncology.

1. Introduction

Brain tumor diagnosis from Magnetic Resonance Imaging (MRI) is one of the most consequential tasks in clinical neuro-oncology. MRI provides non-invasive, high-resolution structural information that enables the differentiation of major intracranial tumor types — glioma, meningioma, pituitary tumor, and no-tumor — each of which demands a distinct clinical management pathway. Manual interpretation of brain MRI by radiologists is time-intensive, subject to inter-rater variability, and increasingly constrained by the global shortage of specialist expertise [4]. These pressures have motivated substantial research into automated deep learning systems capable of supporting radiological decision-making at scale.

Despite the clinical urgency of this problem, existing automated brain tumor classification systems face three interconnected limitations that restrict their real-world utility.Initially, most methods make use of two-dimensional feature extraction from individual MRIs based on a single MRI slice, considering each slice as independent and thus ignoring the intrinsic ordering information between subsequent slices [40, 31]. Second, temporal modeling — when incorporated — is typically implemented through recurrent architectures such as LSTMs or GRUs, which are sequentially constrained, computationally expensive to optimize, and prone to vanishing gradient issues on long sequences [23]. Third, and most critically for clinical adoption, the majority of high-performing models operate as black boxes, offering no mechanism by which clinicians can verify the anatomical basis of a prediction. A model that achieves high accuracy without interpretability cannot be safely integrated into clinical workflows where accountability and auditability are regulatory requirements.

Beyond standalone classification, systems of this kind are increasingly positioned within the broader paradigm of computer-aided diagnosis (CAD), where the objective is not to autonomously diagnose but to augment radiologist decision-making with a consistent, reproducible second opinion. A CAD-oriented brain tumor classification system must satisfy stricter requirements than a benchmark classifier: it must generalize across acquisition protocols, remain robust to inter-patient variability, and — critically — expose its reasoning in a form that a radiologist can audit against established anatomical and radiological criteria. The Swin-TCN-XAI framework is designed with this CAD role in mind, combining high classification performance with a multi-level explainability pipeline intended to support, rather than obscure, clinical judgment.

Prior work has addressed parts of these limitations in isolation. Although CNNs have provided good baselines for MRI classification, they suffer from an intrinsic restriction to localized receptive fields, which makes them unsuitable for modeling long-range spatial relationships within the images [35, 46]. Vision Transformers (ViTs) and their hierarchical variants address this by modeling global attention, yet most ViT-based approaches still process slices independently without modeling inter-slice relationships [18]. Hybrid ViT-GRU models have combined spatial transformers with recurrent sequential modeling and demonstrated strong performance with post-hoc explainability tools such

ORCID(s):

as SHAP and LIME [2, 5]; however, these approaches rely on recurrent architectures that introduce optimization instability and do not benefit from the parallelism and stable gradient properties of convolutional sequential models. Furthermore, the explainability strategies employed in prior work are predominantly post-hoc and single-modality, providing either spatial heatmaps or feature rankings but rarely both simultaneously in a clinically integrated framework [1, 9].

To address these interconnected limitations, the present work introduces **Swin-TCN-XAI**, a hybrid end-to-end deep learning framework for multi-class brain tumor MRI classification. The system combines a pretrained Swin Transformer Tiny backbone with a Temporal Convolutional Network (TCN) classification head, jointly optimized through a differential learning rate strategy. The Swin Transformer extracts hierarchical spatial patch embeddings from each MRI slice through its shifted window self-attention mechanism, producing a 7×7 token grid of dimension 768. The TCN then models ordered inter-token dependencies across this sequence through stacked dilated causal convolutions with residual connections, capturing structured spatial relationships that single-slice CNN or ViT approaches overlook. Joint end-to-end fine-tuning — applying a conservative learning rate of 2×10^{-5} to the pretrained backbone and a faster rate of 3×10^{-4} to the TCN head — enables domain adaptation without degrading ImageNet-learned representations.

The system further incorporates a three-level explainability module: Grad-CAM generates class-discriminative spatial heatmaps by backpropagating through the final TCN convolutional block; SHAP (via DeepExplainer) quantifies the marginal contribution of each Swin token to the classification decision at both global and class-wise levels; and LIME provides instance-level superpixel attributions for individual MRI slices. The combination of these three complementary methods ensures that the framework's predictions can be verified against known anatomical landmarks from spatial, feature-level, and instance-level perspectives simultaneously.

The framework is evaluated on a brain MRI dataset comprising 9,291 training and 1,857 test samples across four classes, under a strict patient-wise train-test split that prevents data leakage by ensuring no MRI slices from the same patient appear in both partitions. The system achieves a classification accuracy of **99.41**%, with macro-averaged precision, recall, and F1-score of 0.9940, 0.9948, and 0.9944, respectively, outperforming all five evaluated temporal and hybrid baselines on every reported metric. The notumor class achieves perfect precision and recall (1.0000), and the glioma class — the most challenging tumor type — attains an F1-score of 0.9925. Convergence is achieved at epoch 17 out of 30, with validation accuracy exceeding 97% from epoch 5 onwards.

The major contributions of this work are as follows:

- We propose Swin-TCN-XAI, a hybrid end-to-end framework that jointly fine-tunes a Swin Transformer and a Temporal Convolutional Network (TCN) to learn complementary spatial and inter-token representations for multi-class brain tumor MRI classification.
- We incorporate a multi-level explainability framework combining Grad-CAM, SHAP, and LIME to provide spatial, global, and instance-level interpretations, enhancing the transparency of model predictions.
- We validate the framework under a strict patient-wise train-test protocol, achieving 99.41% accuracy and outperforming state-of-the-art transformer, temporal, and hybrid baseline models.

This paper is further structured as follows. In Section 2, we introduce related work in the fields of brain tumor classification based on MR images, hybrid DNNs, and XAI. The dataset and the Swin-TCN-XAI framework used are described in Section 3. Experimental results and their quantitative analysis are presented in Section 4. Clinical implications of our findings are discussed in Section 5. Possible limitations and future work are explored in Section 6. Finally, Section

2. Literature Review

Significant advancements in MRI-based brain tumor classification models can be attributed to the development of hybrid models that combine Convolutional Neural Networks (CNN), Vision Transformers (ViT), and Explainable Artificial Intelligence (XAI) techniques. Although the primitive models based on CNN provided robust baselines for the detection of brain tumors, they suffered from limited receptive fields, which made it difficult for these models to incorporate global perspectives. Thus, models based on transformers and hybrid models that combine CNN and transformers have lately gathered popularity owing to improvements in representational capabilities and the development of interpretability techniques.

The model proposed by Ahmed et al. [2] was a hybrid model that used Vision Transformer (ViT) and Gated Recurrent Unit (GRU) models, along with multiple explainable AI (XAI) methods such as attention map, SHAP, and LIME, which resulted in a high level of accuracy of 98.97% on BrMRI-2023 and Kaggle MRI datasets. Interestingly, sequential modeling can be very effective when combined with transformer encoders that capture spatial features in images. Karagoz et al. [18] proposed a self-supervised Residual Vision Transformer (ResViT) that performed very well in a low-data setting, which clearly reveals that using residual or GRU components can be very helpful in coping with some deficiencies in solely CNN-based models.

Although transformers are successful, CNN transfer learning techniques are still competitive. Reddy et al. [35] proposed BrainCD-Net, which employed a DenseNet-169 based feature extractor and traditional classifiers, yielding an accuracy of 95.10%. Similarly, Zhao et al. [46] demonstrated that a ResNet-50 based fine-tuned network can also provide high levels of classification accuracy on brain MRI

datasets, thereby proving that an adequately optimized CNN architecture still has its importance.

Explanation techniques have emerged as the need for explainability in clinical applications. Aly et al. [5] integrated ViT, GRU, SHAP, and LIME techniques for enhanced explainability, demonstrating high accuracy on BraTS benchmarks. While these downstream explainability techniques are beneficial, they may often lack robustness in clinical applications, which could be biased. As an entirely different approach, Sharma et al. [1] improved CNN models by applying attention mechanisms at the channel level, enhancing robustness in noisy environments, but, as known, this is restricted to spatial features.

Recent hybrid approaches have also incorporated multi-branch and ensemble approaches. In a combined ensemble of VGG16 and ResNet50, Dhakshnamurthy et al. [9] achieved a highly accurate classification result. However, a lack of analysis of explainability, volumetric reasoning, and validation by clinical experts is a major issue in these approaches.

While these studies have demonstrated promising performance, several limitations remain. Most existing approaches primarily focus on spatial feature extraction from individual MRI images, with limited exploration of structured inter-token dependencies that capture complementary contextual information. In addition, many explainability methods rely on coarse activation or attention maps, providing only partial insight into model predictions. Furthermore, model evaluation is often conducted on curated benchmark datasets without rigorous patient-wise validation, potentially leading to optimistic performance estimates.

To address these limitations, we propose Swin-TCN-XAI, a hybrid end-to-end framework that combines a Swin Transformer for hierarchical spatial representation learning with a Temporal Convolutional Network (TCN) for modeling structured inter-token dependencies. To enhance clinical interpretability, the framework integrates a multi-level explainability strategy using Grad-CAM, SHAP, and LIME, providing complementary spatial, global, and instance-level explanations. The proposed framework is evaluated using a strict patient-wise train-test protocol on a four-class brain MRI dataset, demonstrating accurate, robust, and clinically interpretable brain tumor classification.

As summarized in Table 1, existing approaches achieve high classification accuracy but often lack temporal modeling, robust interpretability, or evaluation on heterogeneous clinical data.

3. Materials and Methods

This study introduces a hybrid model, referred to as Swin-TCN, which combines Swin Transformer and Temporal Convolutional Networks. A comparative study is also performed against the current state-of-the-art models for deep learning techniques. The aim is to achieve high accuracy in classification while ensuring interpretability by applying Explainable Artificial Intelligence. Preprocessing, training of the model, evaluation, and analysis of explainability make up the workflow (Fig. 1).

3.1. Data Acquisition

The dataset is composed of multiple Magnetic Resonance Imaging (MRI) scans of varying individuals, with the class distribution balanced across glioma, meningioma, pituitary tumor, and non-tumor classes. Realistic inter-center variability is incorporated in the dataset, reducing the bias in the current dataset, considering the differences in image acquisition techniques between centers. Moreover, the images are converted to the grayscale modality to make the representation identical for the neural network, which also reduces the computational complexities. In this task, all the images are converted to the grayscale modality [26]. Sample MRI slices from different tumor categories are illustrated in Fig. 2.

Handling of Missing Values The brain MRI dataset used in this research has already included fully annotated image-label pairs, and there were no missing values within the class labels. Since this dataset is composed only of imaging data, there are no attributes that could be used for imputation or missing value handling. All images used within the experiments were already included before preprocessing, satisfying the quality and completeness criteria.

Statistical Dataset Characteristics To evaluate the characteristics of this dataset, a statistical analysis has been performed. The class-wise sample count has been represented in Table 2. Pixel intensity values are then normalized to the range $[0, 1]$ as part of the preprocessing step. The mean intensity level for all the pixel values was found to be around 0.46, while the standard deviation for all pixel values was calculated to be around 0.21 on average. Spatial consistency is ensured by resampling all the images to a fixed resolution of 224×224 pixels, thus taking care of any variations in spatial resolution that might be present in the MRI slices.

3.2. Preprocessing and Data Augmentation

In order to maintain uniformity across the whole dataset, all the MRI slices are resampled to a fixed resolution of 224×224 pixels. The pixel intensities are normalized to the interval $[-1, 1]$ using a channel-wise mean and standard deviation of 0.5. This ensures that the inputs to the backbone have a stable distribution. No further contrast enhancement steps are taken at the preprocessing stage since the pretrained Swin backbone is adequate enough to account for natural variations in intensity.

The goal of using data augmentation is to improve the generalization process while preventing overfitting. In this way, the data augmentation technique is done only during the training stage. The data augmentation techniques include horizontal and vertical flip, rotation up to an angle of $15°$, and color jittering through the use of perturbation in both brightness and contrast [19] with intensity of 0.2. None of the data augmentation procedures is done during evaluation. Some of the augmented MRI slices are illustrated in Fig. 3.

Table 1
Summary of prior studies on brain tumor classification using CNN-, ViT-, and hybrid-based deep models.

No.	Methodology	Key Contribution	Dataset	Identified Limitation
1	Hybrid ViT–GRU with XAI [2]	High accuracy with sequential modeling and post-hoc explainability	BrMRI-2023, Kaggle MRI	Limited volumetric reasoning; dataset bias; post-hoc XAI
2	Self-supervised ResViT [18]	Effective transformer learning under limited data	BraTS, Figshare	Weak MRI-specific pretext tasks; lacks temporal modeling
3	DenseNet169-based BrainCDNet [35]	Efficient CNN-based hybrid pipeline	MRI datasets	No global context modeling; limited interpretability
4	Fine-tuned ResNet50 [46]	Strong CNN transfer learning baseline	Kaggle MRI	Purely 2D; no explainability or sequential analysis
5	ViT–GRU with SHAP and LIME [5]	Improved interpretability for tumor prediction	BraTS 2013/2015	Older datasets; post-hoc explanations only
6	Channel-attention CNN [1]	Robustness to noise and perturbations	Clinical MRI	CNN-only; lacks global and temporal context
7	Hybrid VGG16–ResNet50 Ensemble [9]	Very high classification accuracy	MRI datasets	No XAI, temporal modeling, or external validation

Table 2
Overview of the Brain MRI Datasets Used in This Study

Attribute	Description
Total Images	11,148 MRI slices
Tumor Classes	Glioma, Pituitary, Meningioma, No Tumor
Class Distribution	Glioma: 3,621; Meningioma: 2,619; No Tumor: 2,334; Pituitary: 2,574
Image Modality	Magnetic Resonance Imaging (MRI)
Original Resolution	Variable (scanner-dependent)
Resized Resolution	224×224 pixels
Train/Test Split	80% / 20% (patient-wise)
Training Samples	9,291 slices
Testing Samples	1,857 slices
Validation Set	Derived from training set during optimization

3.3. Data Splitting into Training and Testing Sets

The performance and generalization abilities of the suggested deep learning model were analyzed by splitting the brain MRI data set into training and testing sets in an 80/20 ratio. Finally, the training and testing data sets included 9,291 and 1,857 MRI slices, respectively.

Importantly, the dataset was split in a strict **patient-wise** manner to prevent data leakage between the training and testing sets. This ensured that MRI slices belonging to the same patient were exclusively assigned to either the training or the testing subset, but never both simultaneously. Such a patient-level separation provides a more clinically realistic evaluation protocol, as the model is tested only on entirely unseen subjects rather than previously observed patient distributions.

The overall dataset partitioning strategy is illustrated in Fig. 4.

From a clinical reliability standpoint, this distinction is not a mere methodological formality. A model evaluated under random slice-level splitting may appear highly accurate simply because it has implicitly memorized patient-specific scanner artifacts, anatomical idiosyncrasies, or acquisition noise present in slices it has already seen from that same subject. Such a model offers no guarantee of performance on a genuinely new patient — the exact scenario it would face in real clinical use. Patient-wise evaluation removes this confound, so that the reported accuracy reflects generalization to unseen individuals rather than recall of previously encountered scan characteristics.

Prevention of Data Leakage A static patient-level train-test split was strictly followed to ensure that the model is evaluated without bias. This ensures that all the MRI slices of a particular patient were either assigned to the training set or the test set, without any intersection between the two sets. This ensures that the model is evaluated on unseen patients, thereby giving a real-world idea of the performance of the model.

3.4. Image Data Loading via PyTorch ImageFolder

Images are read using `torchvision.datasets.ImageFolder` from PyTorch, which automatically determines the classes based on the folder names (glioma, meningioma, notumor, pituitary). Data batching and loading are accomplished using `torch.utils.data.DataLoader` with a batch size of 32 and 2 workers. The memory-efficient method allows for efficient

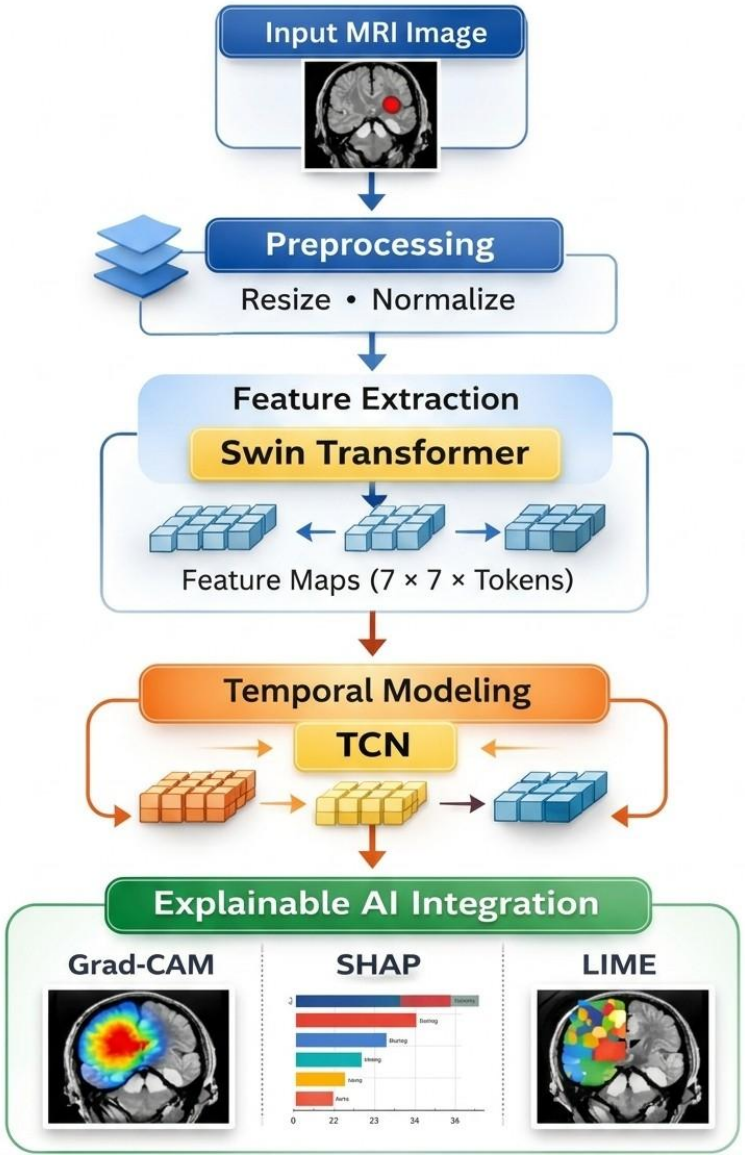


Figure 1: Schematic illustration of the proposed architectural framework.

transformations of images on the fly and is perfectly compatible with the Swin Transformer and TCN modules based on PyTorch [32].

3.5. Label Encoding

Classes are labeled using numerical values (0 for glioma, 1 for meningioma, 2 for notumor, and 3 for pituitary), and they are directly fed into nn.CrossEntropyLoss from PyTorch, which takes care of applying the log-softmax. In order to avoid overconfidence and improve calibration, label smoothing is applied with a value of 0.1, as mentioned by [17]. Using this approach, there is no need for the one-hot representation.

3.6. Feature Scaling via Pixel Normalization

MRI pixel intensities are normalized using channel-wise mean and standard deviation of 0.5, mapping the input range from $[0, 255]$ to $[-1, 1]$. This normalization is applied uniformly across all three channels after converting images to tensors via `transforms.ToTensor()`, and is consistent with the input statistics expected by the ImageNet-pretrained Swin Transformer backbone. Normalized inputs ensure stable gradient flow and faster convergence during end-to-end fine-tuning [12].

3.7. Algorithm

For this proposed approach, two key algorithms were employed: Swin Transformer for hierarchical extraction of features and Temporal Convolutional Network (TCN) for sequential representation of features at a patch level. These were combined with Grad-CAM to achieve explainable tumor localization. In addition to that, SHAP (Shapley Additive exPlanations) analysis was used to determine the weight

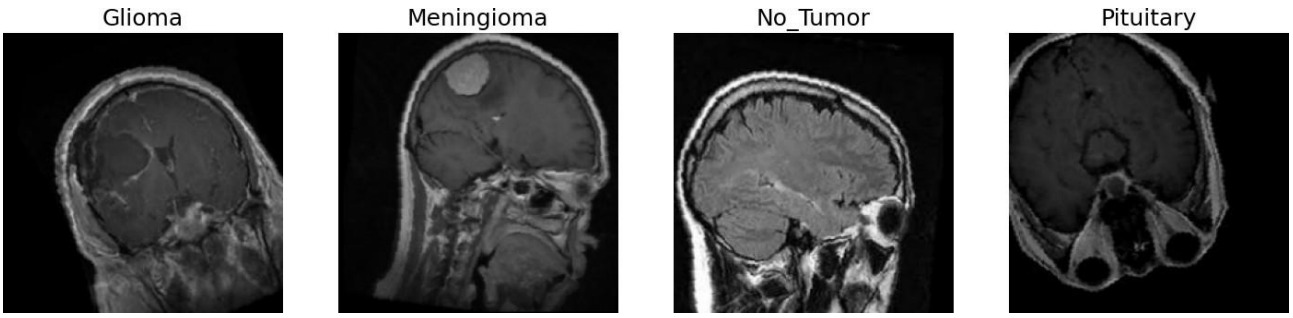


Figure 2: Representative samples drawn from the brain tumor MRI dataset, with emphasis on the different tumor categories.

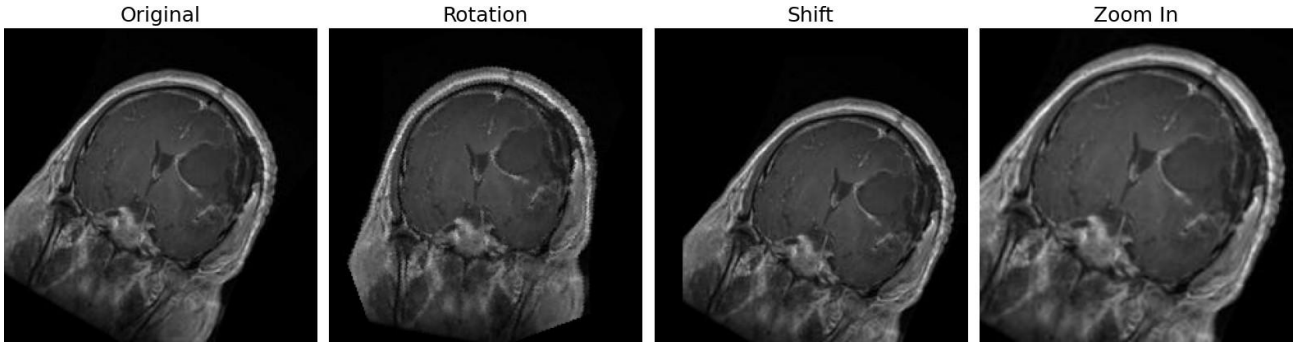


Figure 3: Data augmentation techniques that have been used on the MRI images to show the diversity of the dataset.

that features from the input contribute to reaching a final model prediction.

Justification of Swin Transformer → TCN Fusion Order The fusion order for the proposed Swin-TCN model architecture is set such that the Swin Transformer model comes before the Temporal Convolutional Network (TCN) model. The Swin Transformer acts as a hierarchical spatial feature extractor due to the usage of the shifted window self-attention mechanism, which helps in capturing local as well as global context for the tumor from the input magnetic resonance imaging (MRI) slices. This model produces compact, semantic features that represent the complex spatial features from the input brain MRI images.

The proposed TCN model is then used for capturing ordered inter-slice dependencies from the input MRI scan. This model is more appropriate for capturing such dependencies since TCN models can efficiently capture sequential relationships from input data using dilated convolutions while also ensuring stable gradients during training for parallelization.

The reverse fusion pattern, which is TCN followed by Swin Transformer (TCN → Swin Transformer), is comparatively inefficient and impractical. This is because temporal convolutional networks (TCNs) are designed for modeling

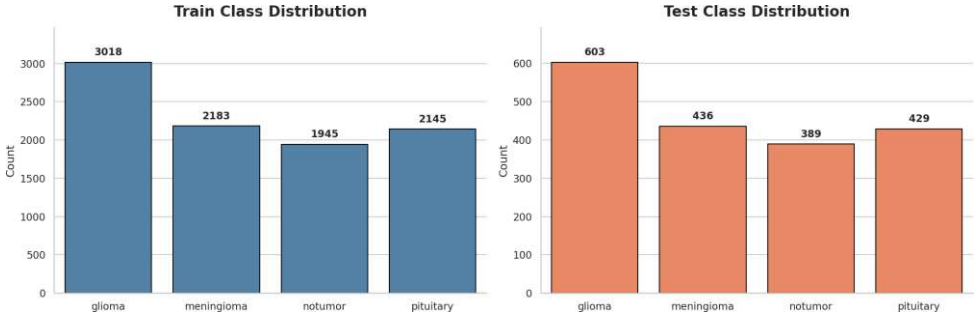


Figure 4: Diagrammatic representation of the partitioning of the dataset into training and testing subsets.

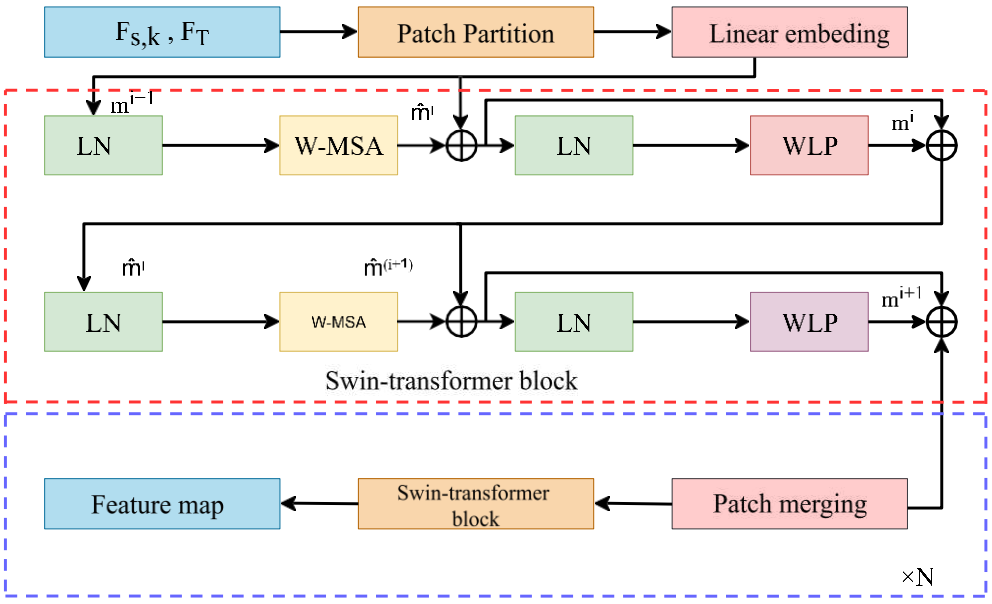


Figure 5: Flow diagram of Swin Transformer.

structured sequential representations of features rather than high-dimensional representations of image pixels. Using TCNs for pixel-level information would significantly increase the computational costs and compromise the ability of the transformer to create meaningful hierarchical representations. This makes the fusion pattern of Swin Transformer followed by TCN (Swin → TCN) computationally efficient and capable of modeling spatial and temporal information.

3.7.1. Swin Transformer

Swin Transformer is a hierarchical vision transformer that incorporates a shifted window attention mechanism to effectively encode local and global information. The multi-scale structure of Swin Transformer allows patch embeddings to be processed at different scales, thereby aiding in effective tumor feature extraction. Figure 5 illustrates the Swin Transformer workflow adopted in this study.

The Swin Transformer model proposed here shows its efficiency in extracting local features as well as global features from the input image slices, making it a viable model for tumor classification problems. The process begins with the division of the input image into non-overlapping patches and then embedding these patches as tokens. Further processing takes place through a sequence of hierarchical layers, where the W-MSA module verifies the extraction of local features, and the SW-MSA module helps in the communication between distant windows to share global information. A modified approach proposed in a recent study by Pacal et al. [30] utilized a Swin Transformer with a ResMLP to achieve excellent tumor diagnosis results in the context of multiple MRI tumor classes. A joint architecture of Swin Transformer + ResNet50V2 was also used in a study, which minimized computation complexity while maximizing accuracy performance in both binary and four-class MRI datasets [45]. These modified models have validated the effectiveness of the suggested architecture in the efficient classification of tumors through brain MRI data. The patch merging layers reduce the spatial resolution of the feature map as the depth of the channel increases, resulting in the creation of a deep multi-scale feature representation.

To explain this procedure in an open way step-by-step, **Algorithm 1 illustrates the Swin Transformer feature extraction used here**.

Patch Merging Downsampling feature maps to build a hierarchical representation:

$$\mathbf{Z}_{s+1} = \text{Proj}\left(\text{Concat}_s\left(\mathbf{Z}_{(2\times 2)}\right)\right), \quad (1)$$

which constructs a feature pyramid, later interfacing with the TCN module.

Working Mechanism Swin Transformer splits a feature map into M × M windows to conduct local attention, and cyclic shifts help windows communicate through linear computational complexity. Swin Transformer is particularly appropriate for cancer classification tasks, which require both fine-grained and global contextual information.

Patch Partitioning and Embedding Input slices $\mathbf{X} \in \mathbb{R}^{H\times W\times 1}$ are tokenized via non-overlapping $P \times P$ patch embeddings, yielding:

$$\mathbf{Z}_0 \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times C} \quad (2)$$

Hierarchical Stages The architecture is divided into different phases, which combine Window Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA). The patch merge module reduces the image resolution by half and doubles the number of

Algorithm 1 Swin Transformer Feature Extraction for Brain Tumor Classification

Initialization:
X: Input MRI slice of size $H \times W \times 1$
θ: Hyperparameters (Table 3)
P: Patch size (4×4), M: Window size 7, S: Swin stages 5:
£ = {Precision, Recall, F1-Score, Accuracy}

procedure DATAPREPARATION
  Normalize **X** to $[0, 1]$
  Partition **X** into non-overlapping patches of size $P \times P$
  Embed each patch to obtain token embeddings $\mathbf{Z}_0$
end procedure

procedure MODELTRAINING
  for $s = 1$ **to** S **do**
    Apply Window Multi-Head Self-Attention (W-MSA)
    Apply Shifted-Window MSA (SW-MSA) for cross-window communication
    Perform patch merging (reduce spatial resolution, double channels)
  end for
  Obtain final hierarchical feature map $\mathbf{Z}_S$
end procedure

procedure PERFORMANCEEVALUATION
  Use extracted $\mathbf{Z}_S$ for downstream classification
  Compute £ = {Precision, Recall, F1-Score, Accuracy}
  Select best-performing attention configuration
end procedure

Main Execution:
DATAPREPARATION
MODELTRAINING
PERFORMANCEEVALUATION
return Swin Transformer feature map $\mathbf{Z}_S$ and evaluation results

channels, making it suitable for tumor analysis at different scales [22].

Relative Position Bias A relative positional encoding indexed by token offsets is employed by each attention head. This makes them invariant to location. Efficiency is improved with Flash Window Attention.

Task Heads Global pooling layers are followed by fully connected layers to derive the final predictions, making it easy to incorporate with hybrid pipelines like Swin-TCN-XAI [14].

Mathematical Formulation **Window Attention (W-MSA).** Window Attention (W-MSA). The self-attention between each non-overlapping window is written as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}} + B\right)V, \quad (3)$$

where B denotes the relative position bias [43].

Shifted-Window Attention (SW-MSA). To enable cross-window communication, a cyclic shift of $(\lfloor M/2 \rfloor, \lfloor M/2 \rfloor)$ is added prior to W-MSA, weaving in with it to capture local and global context. [41].

3.7.2. Temporal Convolutional Network (TCN)

Each MRI scan is represented as an ordered sequence of axial slices, where Swin Transformer extracts patch-level features per slice, and the TCN models ordered inter-slice dependencies across the scan. A temporal convolutional network (TCN) models the relationship between sequences by employing dilated convolutions. As opposed to recurrent models, TCNs are able to train in parallel yet still model both short and long-term dependencies effectively. The temporal dependencies, in this case, represent the sequence through which the images are arranged in order as consecutive MR scans, and not through time series imaging. This approach is able to capture the contextual relationship between the images, specifically the ordered images within the scan, allowing the TCN to model these relations effectively . Fig. 6 illustrates the Temporal Convolutional Network (TCN) process applied in this study.

To provide a brief, step-by-step demonstration of this temporal feature extraction process, **Algorithm 2 illustrates the TCN feature extraction used within this research**.

Mathematical Formulation For an input sequence $\{z_1, \ldots, z_n\}$, the TCN output at time t is:

$$y_t = \sum_{k=0}^{K-1} w_k \cdot z_{t-d \cdot k} \quad (4)$$

where K is the kernel size, d is the dilation factor, and w_k are learnable weights. Stacking dilated layers expands the receptive field exponentially, while residual connections improve training stability.

Architecture

- **Input:** Flattened patch embeddings (from Swin Transformer) serve as the input sequence [10].
- **Dilated Convolutions:** Successive dilations ($d = 1, 2, 4, 8, \ldots$) capture progressively longer dependencies.
- **Residual Blocks:** Each block integrates convolution, normalization and ReLU activation to ensure stability [24].
- **Output:** Global pooling followed by dense layers generates the final class predictions.

3.7.3. Swin–TCN–XAI Fusion for Brain Tumor Classification

In the Swin-TCN fusion, the spatial extraction of features offered by the Swin Transformer is combined with the temporal capabilities of the Temporal Convolutional Network. In this fusion, the features extracted from patches by the Swin Transformer are flattened, which are then processed by the TCN to extract temporal information. **Algorithm 3** encapsulates this hybrid pipeline with feature extraction, fusion, classification, and explainability.

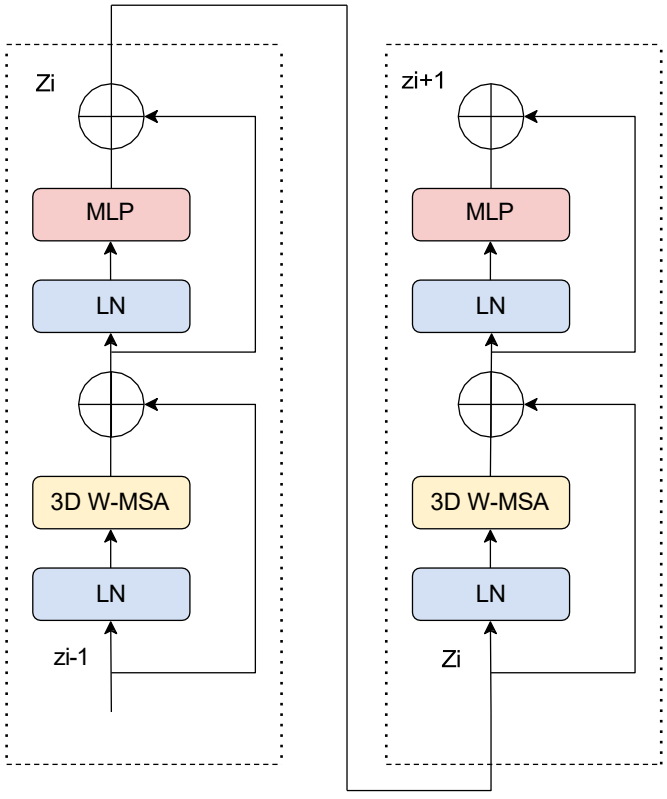


Figure 6: Flow diagram of TCN.

Algorithm 2 Temporal Convolutional Network (TCN) for Sequential Feature Modeling

Initialization:
Input feature sequence $\mathbf{Z} = \{z_1, z_2, \ldots, z_n\}$ from Swin Transformer
θ: Hyperparameters (Table 3)
K: Kernel size = 3, dilations $\{1, 2\}$, residual blocks = 2
£ = {Precision, Recall, F1-Score, Accuracy}

procedure DATAPREPARATION
    Flatten $\mathbf{Z}_S$ into sequential representation of length 49
    Normalize features and prepare batches for input to TCN
end procedure

procedure MODELTRAINING
    for each residual block **do**
        Apply 1D dilated convolution: $y_t^{(l)} = \sum_{k=0}^{K-1} w_k^{(l)} z_{t-k}^{(l-1)}$
        Apply ReLU activation and Layer Normalization
        Add residual connection: $z_t^{(l)} = y_t^{(l)} + z_t^{(l-1)}$
    end for
    Apply global average pooling followed by fully connected layers
    Optimize using AdamW with lr = 3×10^{-4} and cosine annealing scheduler
end procedure

procedure PERFORMANCEEVALUATION
    Evaluate predictions $\hat{y}$ on test data
    Compute £ = {Precision, Recall, F1-Score, Accuracy}
    Select best-performing dilation and kernel settings
end procedure

Main Execution:
DATAPREPARATION
MODELTRAINING
PERFORMANCEEVALUATION
return Temporal feature representation and performance metrics

Algorithm 3 Hybrid Swin–TCN–XAI Framework for Brain Tumor Classification

Initialization:
X: Input MRI slice, θ: Hyperparameters (Table 3)
Backbone lr = 2×10^{-5}, Head lr = 3×10^{-4}, epochs = 30
£ = {Precision, Recall, F1-Score, Accuracy}

procedure DATAPREPARATION
    Normalize MRI slices and split dataset into train/test
    Apply patch extraction and embedding as per Algorithm 1
end procedure

procedure MODELTRAINING
    Stage 1: Swin Transformer Encoding
    Extract hierarchical features $\mathbf{Z}_S$ using shifted-window attention
    Stage 2: Temporal Modeling with TCN
    Pass $\mathbf{Z}_S$ to TCN with dilated convolutions and residual blocks
    Stage 3: Classification Head
    Apply global pooling → fully connected → Softmax
    Optimize with AdamW (differential learning rates) and cosine annealing scheduler
end procedure

procedure PERFORMANCEEVALUATION
    Predict tumor class $\hat{y}$ on test set
    Compute £ = {Precision, Recall, F1-Score, Accuracy}
    Explainability Stage:

    Generate Grad-CAM, SHAP, and LIME explanations after the classification stage.
end procedure

Main Execution:
DATAPREPARATION
MODELTRAINING
PERFORMANCEEVALUATION
return ($\hat{y}$, Grad-CAM, SHAP, Lime, £)

Rationale and Design Goals However, the proposed hybrid framework uses the strengths of Swin Transformer, Temporal Convolutional Network (TCN), and Explainable Artificial Intelligence (XAI) to address brain tumor classification in MRIs. Swin Transformer focuses on hierarchical spatial patterns, while TCN modules address temporal dependencies. On the other hand, Grad-CAM with SHAP handles interpretability [25]. This design ensures not only high accuracy but also clinical transparency.

Hierarchical Feature Encoding via Swin Transformer To attain the stated goals, the Swin Transformer Network makes use of the shifted window self-attention approach, thus enabling the capturing of both the local architectural design as well as global information. This is made possible through the hierarchical structure of the Swin Transformation Network, where resolution is reduced through augmentation of semantic representations. Augmentation of the generalization capabilities of the model is achieved through the use of pre-trained weights, as seen in the Swin-SMT.

Token Sequencing and Swin–TCN Interface To bridge the gap, the outputs from the Swin Transformer are flattened into token form, which are then passed to the Temporal Convolutional Network (TCN). In the TCN, long-distance information within the data is modeled using dilated causal convolutions, such that the fusion features,**F**, include the information related to the morphological characteristics associated with the local regions.

Temporal Convolutional Network (TCN) Head This Temporal Convolutional Network (TCN) head, made up of dilation convolutions, normalizations, and nonlinear activations, is used to provide regularization to the training phase. Pooling, followed by fully connected layers, provide the final output classes in terms of $\hat{y} = \text{Softmax}(W\ \mathbf{F} + b)$. Tempo-ral Convolutional Networks address the complex temporal aspects in medical image analysis, hence supporting the proposed method [37].

Training and Computational Efficiency The AdamW optimizer is employed with differential learning rates: 2×10^{-5} for the pretrained Swin backbone and 3×10^{-4} for the TCN classification head. A cosine annealing learning rate scheduler reduces the learning rate to 10^{-6} over 30 training epochs. Cross-entropy loss with label smoothing of 0.1, dropout regularization at rate 0.3, weight decay of 10^{-4}, and gradient clipping with $\max_norm = 1.0$ collectively ensure stable convergence and strong generalization. The window-based self-attention in the Swin Transformer and the parallelizable dilated convolutions in the TCN ensure that the end-to-end model remains computationally tractable on modern GPU hardware.

Overfitting Control and Regularization Strategies Several different forms of regularization were used to minimize overfitting while improving the generalizability of the model. Dropout was used in several layers of both the Swin Transformer and TCN models. Through the process of randomly turning off the activation of neurons in the training phase, this form of regularization prevented features from co-adapting.

Furthermore, a number of data augmentation methods such as flipping, rotating, and changing image scale and intensity were used to mimic real-world imaging situations. Lastly, early stopping was used to stop training once the performance on validation measures began stagnating. Taken together, these strategies effectively reduced overfitting while still allowing for successful model convergence.

3.8. Explainability Module (Grad-CAM, SHAP, LIME)

Although the proposed Swin-TCN model has strong predictive power, the application of this model in practical healthcare settings should be clear and interpretable. It has been established that deep neural networks are generally considered to be a “black box.” This can cause some issues concerning trust, especially in healthcare environments. To overcome such difficulties, the proposed work utilizes the concept of multi-level explainability through the use of **Grad-CAM**, **SHAP**, and **LIME**.

In a diagnostic setting, a classification label alone carries limited clinical value unless the radiologist can verify why the model arrived at that decision. Explainability, in this context, functions less as a diagnostic auxiliary and more as a decision-support mechanism: it allows a clinician to cross-check the model’s attention against known tumor morphology before a result is acted upon, rather than accepting the prediction as an opaque output. This is particularly critical in borderline or atypical presentations, where a radiologist may wish to confirm that the model’s decision is anchored to the lesion itself rather than to incidental scan artifacts, motion blur, or acquisition-specific noise.

Thus, it is ensured that the Swin-TCN-XAI model achieves a high level of classification accuracy while offering a meaningful explanation consistent with radiologic interpretation.

3.8.1. Spatial Explainability using Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to the final convolutional block of the Temporal Convolutional Network (`tcn.network[-1]`), which is the last feature-producing layer before the classification head. Forward and backward hooks are registered on this layer to capture activations and gradients during inference. Class-specific importance weights are computed by globally averaging the gradients with respect to each feature map channel, and the resulting weighted activation sum is passed through a ReLU function to produce a class-discriminative heatmap. The heatmap is subsequently reshaped from the 49-token sequence back to the 7×7 spatial grid and upsampled to 224×224 pixels for overlay on the original MRI slice [33].

For a predicted class c, the importance weights are computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

and the corresponding localization map is given by:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{6}$$

These heatmaps allow clinicians to visually verify whether the model focuses on anatomically plausible tumor regions rather than irrelevant background structures.

3.8.2. Global Feature Attribution using SHAP

In order to better understand the explanatory space, the final classifier's feature contribution is quantified by SHAP (SHapley Additive exPlanations). SHAP values, which are derived from cooperative game theory, quantify the contribution of features.

In this paper, SHAP values are utilized for the Swin-TCN model's fused feature embeddings to determine which features contribute the most to the classification of the tumor. The mean of the absolute SHAP values can offer insights into which features mainly contribute to the correct classification of the tumor [16].

3.8.3. Local Instance-Level Interpretability using LIME

To clarify individual MRI slice instance-level explanations, the approach called Local Interpretable Model-Agnostic Explanations (LIME) is used. LIME works by perturbing the confined region around the input image using superpixel segmentation. A sparse linear model is also used around the instance being explained [44].

This results in localized explanations that can be understood by humans since they point out the parts within images that contribute positively or negatively to the outputs that predict the class of the tumor. Thus, LIME fills the gap that exists within the interpretability abilities of Grad-CAM.

The combination of the outputs of the Grad-CAM, SHAP, and LIME techniques helps to enable clinical experts to verify that the model focuses on anatomically valid areas of the tumor and to understand the value of individual features for each of the predictions.

The integration of Grad-CAM, SHAP, and LIME offers a robust explainability model that offers localization, explanation, and elucidation of decision-making processes, and this makes the proposed Swin-TCN-XAI model suitable for transparent and trustworthy deployment in clinical settings.

3.9. Evaluation Metrics

In order to determine the clinical reliability and efficiency of Swin-TCN-XAI framework classification, various metrics for evaluating were considered. These metrics included Precision, F1-Score, Sensitivity (Recall) and Overall Accuracy. These parameters are considered important for medical image classification because they can perform class-wise as well as overall classification, which is useful for class-imbalanced scenarios [19, 21].

3.9.1. Precision

Precision calculates the proportion of true positive predictions out of all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

In addition, a high level of precision is required to avoid false positives and unnecessary medical procedures.

3.9.2. Sensitivity (Recall)

Sensitivity measures how well the model can detect actual tumor instances:

$$\text{Sensitivity} = \frac{TP}{FN + TP} \tag{8}$$

This is even more important in oncological tasks, as an incorrectly detected case can invite catastrophic outcomes.

3.9.3. F1-Score

F1-score is the harmonic mean of the precision and sensitivity:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{9}$$

It performs a balanced evaluation of the models' performance, especially in cases with moderate class imbalance.

3.9.4. Overall Accuracy

Overall accuracy mirrors the number of correctly classified samples in each of the classes:

$$\text{Accuracy} = \frac{TP + TN}{FN + TN + FP + TP} \tag{10}$$

Even if accuracy is not representative of any class differences, it is a global index of reliability for classification.

Taken together, these measures provide an integrated, highly actionable evaluation of performance and generalization capacity for the framework.

3.10. Hyperparameter Configuration

The Swin-TCN-XAI model was trained under a unified end-to-end configuration to ensure stable convergence and strong generalization. All hyperparameter values are summarized in Table 3 and correspond directly to the implementation described in Algorithms 1–3.

The Swin Transformer Tiny backbone (`swin_tiny_patch4_window7_224`) pretrained on ImageNet-1K, serves as the spatial feature extractor. Input MRI slices are divided into non-overlapping 4 × 4 patches with a window size of 7 tokens, producing a

Table 3
Hyperparameters of the proposed Swin-TCN-XAI end-to-end model.

Hyperparameter	Value
Backbone	Swin-Tiny (ImageNet-1K pretrained)
Patch / Window Size	4×4 / 7
Output Feature Map	$7 \times 7 \times 768$ (flattened to 49×768)
TCN Channels	$[128, 256]$
TCN Kernel / Dilations	$3, \{1, 2\}$
Dropout Rate	0.3
Backbone Optimizer	AdamW, lr = 2×10^{-5}, weight decay = 10^{-4}
Head Optimizer	AdamW, lr = 3×10^{-4}, weight decay = 10^{-4}
LR Scheduler	Cosine Annealing ($T_{\max}$ = 30, $\eta_{\min}$ = 10^{-6})
Training Epochs	30 (best checkpoint at epoch 17)
Batch Size	32
Number of Classes	4
Label Smoothing	0.1
Gradient Clipping	max $_norm$ = 1.0
Total Parameters	28,291,582
Explainability (XAI)	Grad-CAM, SHAP, LIME

spatial feature map of size $7 \times 7 \times 768$, which is flattened into a sequence of 49 tokens of dimension 768 before being passed to the TCN module.

The TCN classification head employs dilated causal convolutions with kernel size 3 and dilation rates $\{1, 2\}$, with output channels $[128, 256]$. Dropout at rate 0.3 and weight decay of 10^{-4} are applied for regularization. The backbone and head are jointly optimized using the AdamW optimizer with differential learning rates: 2×10^{-5} for the pretrained Swin backbone and 3×10^{-4} for the TCN head, preventing degradation of ImageNet-learned representations while allowing rapid adaptation of the classification head. A cosine annealing scheduler reduces the learning rate from the initial value to 10^{-6} over 30 epochs. Gradient clipping with max $_norm$ = 1.0 and label smoothing of 0.1 further stabilize training and improve output calibration. The best model checkpoint is saved at epoch 17 based on validation accuracy.

For explainability, Grad-CAM is applied to the final TCN convolutional block via registered forward and backward hooks, producing class-discriminative spatial heatmaps. SHAP (via DeepExplainer) is used to compute token-level feature attribution across the test set, while LIME provides instance-level superpixel explanations for individual MRI slices.

All hyperparameters were kept fixed across experiments to ensure fair comparison and reproducibility.

3.11. Reproducibility and Implementation Availability

In order to ensure full reproducibility of the study, the preprocessing methods, model parameters, and training procedures have been clearly specified. In this study, the Swin Transformer backbone model has been used and initialized using the publicly available pre-trained weights on the ImageNet dataset.

For the training and evaluation of the model, PyTorch 2.x and the `timm` library were used exclusively. The random seed was fixed to 42 across all components (`random`, `numpy`, `torch`) to ensure reproducibility of the results. Experiments have been conducted under the same preprocessing methods, training procedures, and evaluation protocols.

The dataset used in the study is publicly available, and the experiment configurations can be replicated using standard deep learning environments.

4. Results Analysis and Discussion

In this paper, a comprehensive experimental analysis has been performed to evaluate the **Swin-TCN-XAI** framework for multi-class brain tumor analysis performed on magnetic resonance imaging scans. The analysis of the hybrid framework includes training stability, accuracy, discrimination ability, and interpretation capabilities. At the same time, comparative studies are carried out in order to examine the effectiveness and generalizability of the new approach with respect to different models.

All experiments are performed under identical environmental settings for the preprocessing, training, and testing phases. The evaluation is performed based on certain parameters such as **Precision, Sensitivity, F1 Score, and Overall Accuracy**, considering their prior established relevance for medical image classification problems, which are often characterized by class imbalance, robustness, and accuracy. Explainability is another area of interest, and this is measured based on parameters such as **Grad-CAM, SHAP, and LIME**, etc.

In the experimental analysis, the discussion proceeds as follows. First, the convergence behavior of the Swin-TCN network is analyzed through the training and validation losses, as well as the validation accuracy. Second, confusion matrices and classification reports are used. Finally, the issue of clinical interpretability is addressed, along with the results obtained from the Grad-CAM, SHAP, and LIME analyses, as well as the model comparison test.

4.1. Experimental Setup and Implementation Details

All experiments were conducted using the Python programming language with PyTorch 2.x and the `timm` library for Swin Transformer loading. Training was performed on an NVIDIA RTX 4090 GPU (24 GB VRAM) with an Intel i9 processor and 64 GB RAM. The Swin Transformer backbone (`swin_tiny_patch4_window7_224`) was initialized from publicly available ImageNet-1K pretrained weights. MRI

images were resized to 224×224 pixels and normalized using mean = 0.5 and standard deviation = 0.5 per channel, mapping pixel values to the range $[-1, 1]$. Training augmentations included random horizontal and vertical flips, random rotation up to $15°$, and color jitter with brightness and contrast variation of 0.2.

The framework adopts an end-to-end training strategy in which the Swin backbone and the TCN classification head are jointly optimized using differential learning rates. The AdamW optimizer is applied with a backbone learning rate of 2×10^{-5} and a head learning rate of 3×10^{-4}, both with weight decay 10^{-4}. A cosine annealing learning rate scheduler with $T_{\max} = 30$ and $\eta_{\min} = 10^{-6}$ is applied to gradually reduce the learning rate. Gradient clipping with $\max_norm = 1.0$ is used to stabilize training. Cross-entropy loss with label smoothing of 0.1 serves as the training objective. The model was trained for 30 epochs with a batch size of 32, with the best checkpoint saved at epoch 17 based on validation accuracy.

The dataset was divided into training and testing subsets at an 80:20 ratio following a strict patient-wise split, ensuring that all MRI slices from a given patient appear exclusively in either the training or the test set, with no overlap between the two partitions.

4.2. Training Loss and Validation Accuracy

The proposed end-to-end Swin-TCN model is capable of fast and reliable convergence, as shown in Figure 7. The training loss drops considerably from 0.635 to less than 0.370 after epoch 1 due to the powerful representation capability of the ImageNet pre-trained Swin backbone network that needs minimal adjustments for the brain MRI task. Following that, the training loss progressively decreases and converges to a relatively stable level of 0.355 after epoch 15, which confirms the effectiveness of the cosine annealing scheme in controlling the learning-rate decay.

The validation accuracy achieves 97.42% after epoch 5 and is greater than 98% starting from epoch 10 onwards. The best checkpoint is obtained at epoch 17, where the validation accuracy is 99.41%. Some fluctuations observed between epochs 10 and 15, ranging from 96.72% to 98.44%, can be attributed to the stochastic behavior of the mini-batch optimization algorithm and do not suggest instability. The stable and excellent accuracy level after epoch 5 further validates the efficacy of the differential learning rate method in fine-tuning the pre-trained backbone with a conservative rate of 2×10^{-5} and the newly added TCN head with a more aggressive rate of 3×10^{-4}.

4.3. Confusion Matrix, Classification Report, and Visualization

The confusion matrix and classification report provide a detailed analysis of the proposed approach on a class-by-class basis, allowing for an in-depth analysis of true positives, false positives, and false negatives for various classes of cancers.

4.3.1. Confusion Matrix

The confusion matrix shown in Fig. 8 highlights the excellent discriminative power of the proposed method for all four tumor categories. Diagonal dominance is observed, suggesting the accuracy of classifications made by the framework. As regards the glioma category (603 samples), 596 classifications are true, with 5 and 2 cases falsely classified as meningioma and pituitary tumors respectively, which gives a recall rate of 0.9884. For the meningioma category (436 samples), 433 samples are accurately predicted, out of which only 2 and 1 are falsely identified as glioma and pituitary tumors respectively, resulting in a recall value of 0.9931. All samples of the notumor category (389 samples) are correctly classified (recall = 1.0000, precision = 1.0000). In the pituitary category (429 samples), 428 classifications are true, while one is falsely recognized as meningioma, thus having a recall rate of 0.9977.

Misclassifications are mainly between the glioma and meningioma categories, which can be explained through the radiologic similarity in MRIs between these two tumors, especially in terms of intensity values in T1 weighted sequences. No misclassification was found between notumor category and others, which is highly valuable from the clinical point of view.

4.3.2. Explainability with Grad-CAM

The use of Grad-CAM enhances the explainability of the model by generating class-specific heatmaps to show relevant sections of the MR scans. Grad-CAM generates class-specific heatmaps by calculating the importance of activation maps generated using gradients from the convolutional layers [20, 39].

Mathematical Overview Grad-CAM computes the importance weight α_k^c for each feature map A^k as the average of the target class score y^c gradients with respect to the feature map. The heatmap output is a weighted sum of such maps fed into a ReLU function:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k^c A^k\right) \tag{11}$$

This highlights spatial areas most effective to the model's output while suppressing unnecessary areas, so clinicians can verify if predictions align with tumor locations, as presented visually in Fig. 9.

The Grad-CAM visualization plots have consistently highlighted the tumor-related anatomical areas for all tumor classes, which shows that the model is actually focusing on the relevant areas rather than the background areas. Moreover, the SHAP visualization plots also confirm that the discriminative areas on which the model is making its classification decisions are related to the tumor location along with its varying intensities.

4.3.3. Explainability with SHAP

SHAP (SHapley Additive exPlanations) offers explainability at the level of features through the determination of

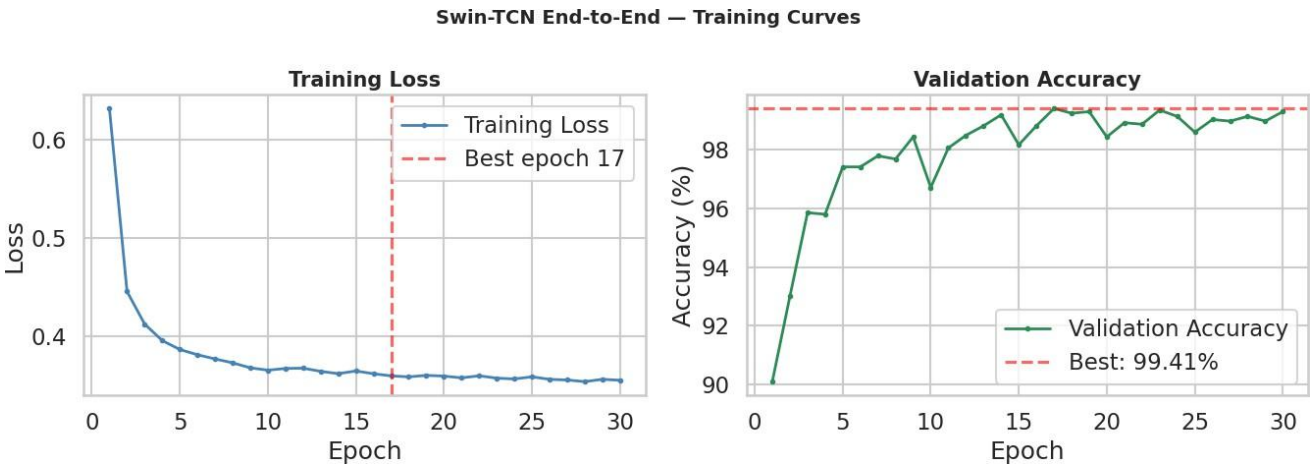


Figure 7: Learning curve for the training loss and validation accuracy of the Swin Transformer-TCN hybrid model.

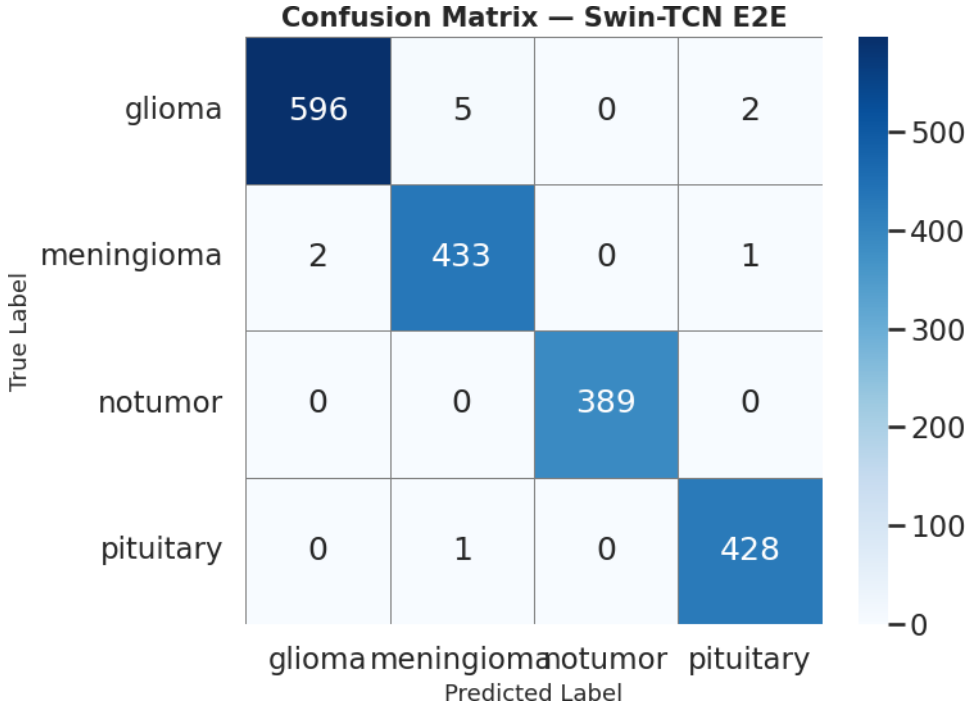


Figure 8: Confusion matrix for the classification results of the proposed Swin-TCN hybrid model.

their contributions to the prediction made by the model. SHAP values, which are based on cooperative game theory, provide a mathematically well-founded way of attributing importance to features in the context of machine learning. Medical imaging studies have found that SHAP can be used to interpret deep learning models applied to brain tumors recognition [15, 29].

Rationale in Medical Imaging In the area of tumor classification using MRIs, SHAP explanation is utilized to ensure that the attributes derived from the model correspond to clinically relevant spatial locations. The combination of Grad-CAM for spatial localization with SHAP for numerical token-based attributions provides a synergistic two-step approach, whereby Grad-CAM highlights the regions where the model focuses its attention, and SHAP computes the amount of influence each token exerts on the final outcome.

Mathematical Foundation of SHAP SHAP computes the Shapley value for feature i as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{12}$$

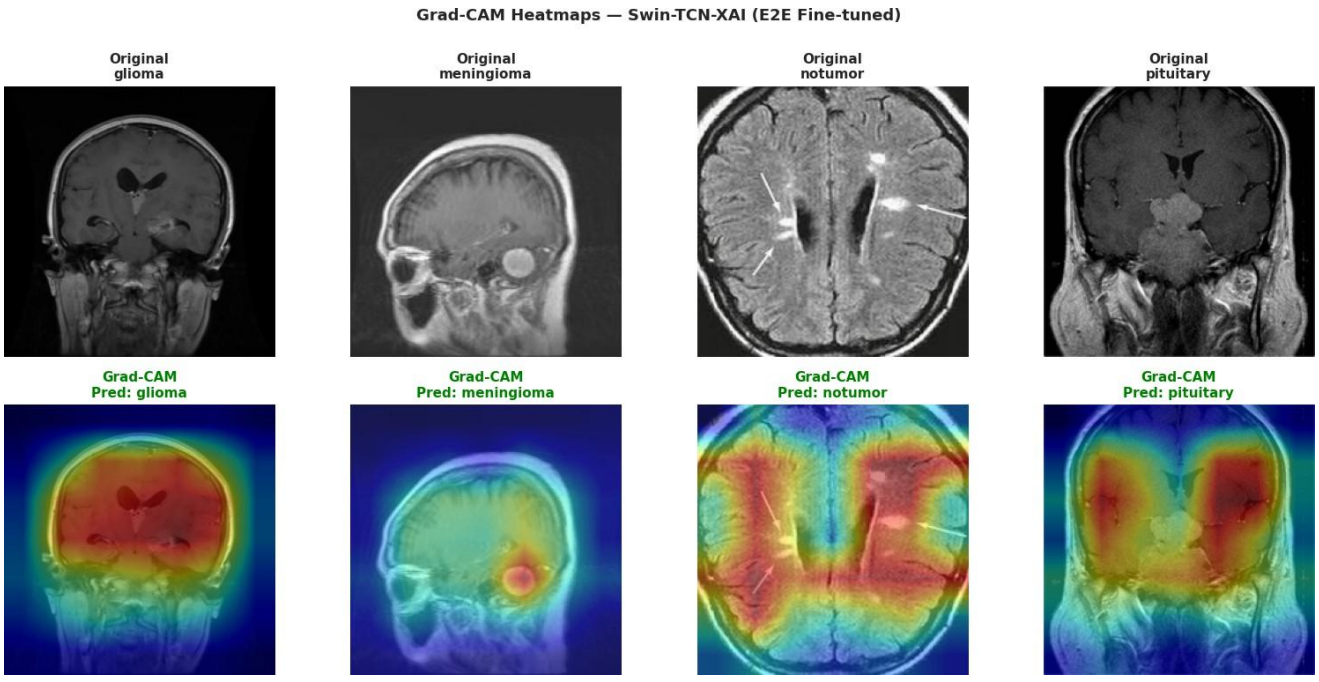


Figure 9: Grad-CAM outputs for different tumor categories.

where F is the complete feature set, S is a subset not containing feature i, and $f(S \cup \{i\}) - f(S)$ denotes the marginal contribution of feature i to the prediction. DeepSHAP, an efficient backpropagation-based approximation, is used for tractable computation over neural network embeddings [7].

Quantitative SHAP Feature Importance The SHAP analysis method was employed for the TCN fused token embeddings on randomly selected 50 test data instances, with a total of 100 background samples drawn from the training dataset. The mean absolute SHAP scores of all test instances were averaged out for the purpose of calculating the global ranking order of token importance. The Swin 7×7 spatial grid of tokens served as the unit of attribution here. The top-10 most important tokens based on mean absolute SHAP scores are listed in Table 4.

The clustering of the high-importance tokens in central locations of the 7 × 7 space is in line with the location of the brain tumors, which are mostly located in central and periventricular areas on MRI images. Periphery tokens (edges of the spatial grid) exhibit low SHAP scores, as they correspond to the skull/scalp areas with low importance for the prediction task.

Class-wise SHAP Attribution Analysis Class-specific SHAP analyses show significance and clinical relevance in terms of attribution patterns. When predicting gliomas, attributions are highest for central/periventricular tokens, in accordance with the diffuse nature and irregular margins of gliomas. On the other hand, when predicting meningiomas, the highest attributions are observed for tokens that are relevant to the borders of the peripheral/cortical regions, due to the extra-axial location and sharp margins of meningiomas. For predictions related to pituitary tumors, attributions are highest for tokens inside the inferior/central grid cells. This is in line with the nature of pituitary tumors which are usually found at the sellar-suprasellar region. Lastly, when predicting the no-tumor class, all attributions are generally small in magnitude.

Table 4
Top-10 Swin Token Importance Ranking by Mean Absolute SHAP Value (computed over 50 test samples)

Rank	Token Position	Mean $\lvert\phi\rvert$
1	Token_16 (centre-left spatial region)	0.0940
2	Token_18 (centre spatial region)	0.0933
3	Token_23 (centre-right spatial region)	0.0904
4	Token_25 (centre spatial region)	0.0903
5	Token_32 (lower-centre region)	0.0887
6	Token_30 (lower-centre region)	0.0884
7	Token_17 (centre spatial region)	0.0835
8	Token_22 (centre spatial region)	0.0790
9	Token_31 (lower-centre region)	0.0780
10	Token_9 (upper-centre region)	0.0774

Integration into Swin–TCN–XAI Pipeline The SHAP method is performed post-classification on the flattened

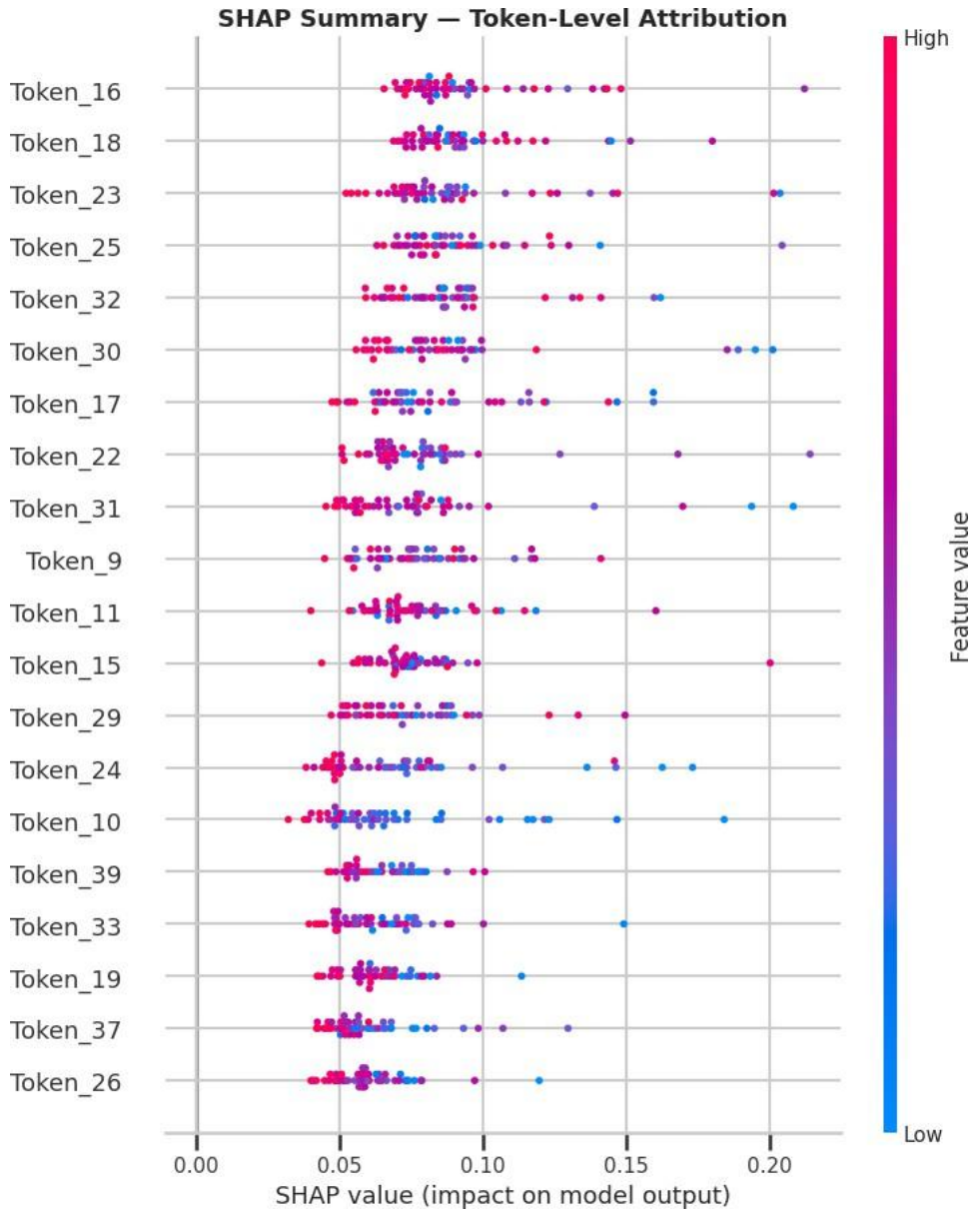


Figure 10: Visualization of the tumor and non-tumor images using the SHAP values.

token embeddings generated by fine-tuning the Swin backbone, using the TCN classifier head as the interpreted function. The summary charts and token maps, as depicted in Fig. 10, generate overall rankings of importance as well as class-specific attribution scores.

4.3.4. Explainability with LIME

LIME (Local Interpretable Model-Agnostic Explanations) provides instance-level transparency by constructing a locally faithful linear surrogate model in the vicinity of each individual prediction. Unlike Grad-CAM, which operates through gradient backpropagation, and SHAP, which computes global token-level attributions, LIME explains individual predictions by perturbing localized superpixel regions of the input image and measuring the resulting change in model output [36]. This makes LIME particularly valuable for per-sample clinical auditing.

In the context of brain MRI classification, LIME partitions each input slice into superpixels using a segmentation algorithm and systematically occludes subsets of these regions to generate perturbed samples. A sparse linear model is then fitted to the model's responses over this neighborhood, assigning positive weights to superpixels that support the predicted class and negative weights to those that suppress it. In the resulting visualizations, positively contributing regions are overlaid in green, while negatively contributing regions are shown in red.

Mathematical Formulation LIME identifies an interpretable explanation model $g \in G$ by minimizing:

$$\xi(x) = \arg\min_{g \in G} \mathcal{G}(f, g, \pi_x) + \Omega(g) \quad (13)$$

where f is the original complex classifier, π_x is a locality-aware proximity kernel centered on the instance x, $\mathcal{G}$ measures the local approximation fidelity, and $\Omega(g)$ penalizes explanation complexity to enforce human interpretability. The formulation balances local faithfulness with explanation simplicity.

Clinical Interpretation Figure 11 shows the LIME explanations obtained for a single image from each of the four different tumor types. In cases where the model predicts glioma, LIME indicates irregular central areas matching the infiltrating nature of gliomas. When the model predicts meningiomas, LIME marks peripheral cortical areas with positive weights, which corresponds to the extra-axial position of meningiomas. Similarly, when predicting pituitary tumors, LIME assigns positive weights to inferior-central superpixels, which corresponds to the sellar position of the pituitary gland. In cases where there is no tumor prediction, LIME assigns positive weights to normal tissue regions having low focal intensity, and none of the lesion-like regions obtain positive weights.

The agreement between LIME superpixel contributions, Grad-CAM spatial attention maps, and SHAP token importance scores for all four classes is compelling proof that the model's choices are always based on anatomically relevant parts of the images and not the background noise of the scans.

4.4. Integrated Interpretation of XAI Results

Integrated explanation for decision-making in the model was approached by concurrently using the SHAP, LIME, and Grad-CAM methods, with the understanding that using only one explanation tool might result in an incomplete or misleading explanation. This combined explanation approach will help in obtaining a holistic, multiple-level explanation for the decision-making mechanism in the Swin-TCN model.

Grad-CAM correctly points to significant anatomical areas for tumors, meaning that it is paying attention to meaningful areas of the input during prediction. SHAP adds another level of meaning by providing the significance of the derived features and embeddings, verifying that the dominant features are those that describe tumor texture and intensity characteristics. Meanwhile, LIME serves as a form of validation for instance-level explanations, segmenting areas that contribute or detract from a prediction.

The fact that there is a high degree of concordance among these three explainability techniques indicates that there are no issues of spurious correlations or background artifacts in the proposed model, because the predictions for it are based on significant clinical structures in global and local views. This helps to reinforce the validity of Swin-TCN-XAI for medical applications.

Table 5
Classification Report for Brain Tumor Detection — Swin-TCN-XAI (E2E)

Class	Precision	Recall	F1-score	Support
Glioma	0.9967	0.9884	0.9925	603
Meningioma	0.9863	0.9931	0.9897	436
No-tumor	1.0000	1.0000	1.0000	389
Pituitary	0.9930	0.9977	0.9953	429
Accuracy	0.9941 (1857 samples)			
Macro Avg.	0.9940	0.9948	0.9944	1857
Weighted Avg.	0.9941	0.9941	0.9941	1857

4.4.1. Classification Report

The classification report, as shown in Table 5, provides a comprehensive per-class evaluation of the proposed Swin-TCN-XAI model on the test set containing 1,857 instances. For the notumor class, the precision and recall are equal to 1.0000, meaning that there is no false positive prediction and false negative prediction of healthy cases, which is a crucial finding from the clinical perspective and justifies the utilization of the framework as a screening tool. For the pituitary class, precision and recall values are equal to 0.9930 and 0.9977, respectively, with an F1-score of 0.9953 over 429 test samples. Precision and recall values for the meningioma class are 0.9863 and 0.9931, respectively, and the associated F1-score is 0.9897 for 436 samples. Finally, for the glioma class, which represents the most challenging radiological case due to its invasive nature and ill-defined boundaries, the precision and recall are equal to 0.9967 and 0.9884, respectively, with an F1-score of 0.9925 for 603 test samples. Overall accuracy for the entire test set comprising 1,857 samples is equal to **99.41**%, with a macro-average precision, recall, and F1-score of 0.9940, 0.

Generalization Performance on External Datasets In order to evaluate the generalization ability of the proposed framework on external data sets, two additional external brain MRI data sets were used. These data sets are the Figshare Brain MRI Data Set [11] and the Mendeley Brain Tumor MRI Data Set (Version 6) [13]. The use of such additional data sets is important since they represent variations in the properties of the MRI scanner, acquisition protocol, image quality, shape and size of tumors, and the distribution of classes. Thus, their consideration helps to understand how well the proposed framework works in other conditions and how much its performance differs from the training conditions.

The results of applying the proposed architecture to external data sets can be found in Table 6. As one can see, the performance of the model remains very high on the Figshare Brain MRI Data Set: its precision is 99.20%, recall is 98.46%, F1-score is 99.44%, and accuracy is 99.11%. Such indicators suggest that the framework shows excellent discrimination ability despite the significant differences between the data set considered and the data set on which

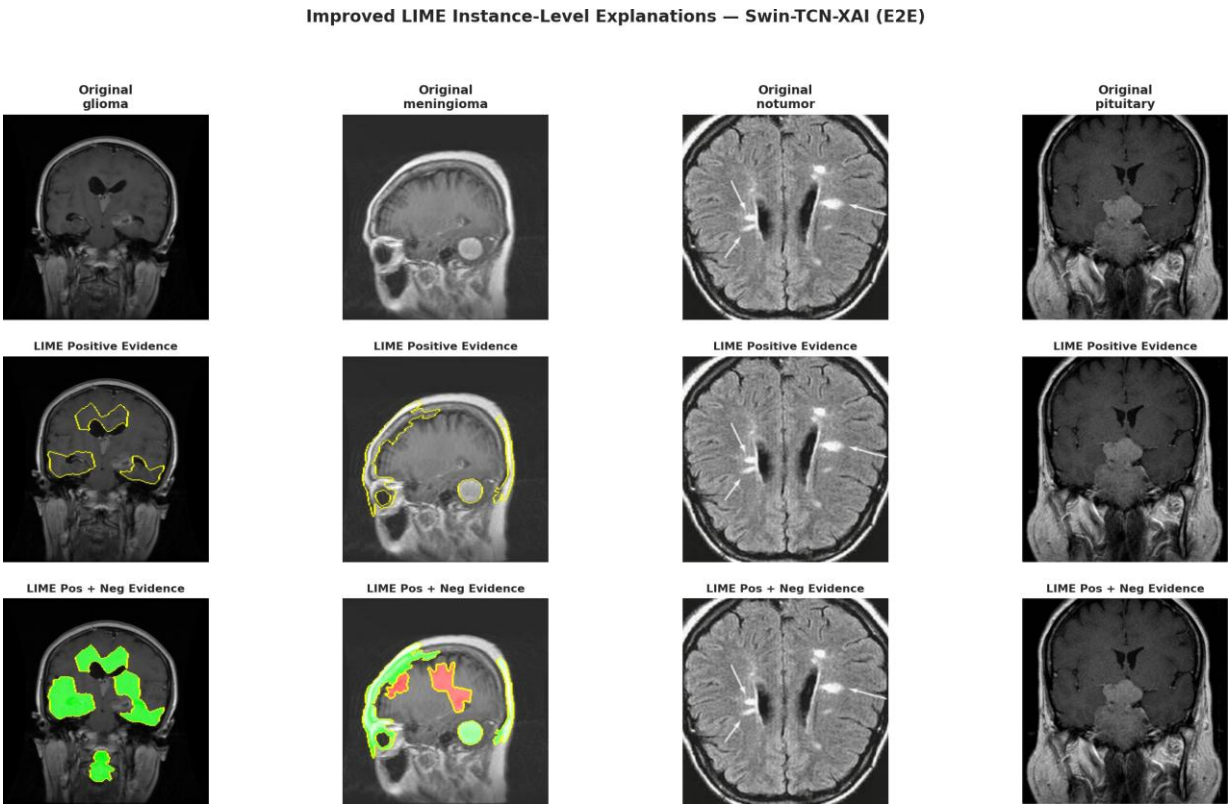


Figure 11: LIME instance-level explanations for representative MRI samples from each tumor class. Green regions indicate superpixels with positive influence on the predicted class; red regions indicate negative influence.

Table 6
Overall classification performance of the proposed Swin-TCN-XAI framework on external brain MRI datasets.

Dataset	Precision	Recall	F1-Score	Accuracy
Figshare MRI Dataset [11]	99.20	98.46	99.44	99.11
Mendeley MRI Dataset [13]	99.38	99.40	99.38	99.41

it was trained. At the same time, on the Mendeley Brain Tumor MRI Data Set, the framework achieved a precision of 99.38%, recall of99.40%, F1-score of 99.38%, and accuracy of 99.41%. The consistency of the results indicates the stability of the architecture under different external conditions.

The high level of performance on external data sets indicates that the hierarchical representations of space obtained using the Swin Transformer are quite transferable to various conditions, and the model is able to maintain high levels of classification performance in any case. In addition, such good results demonstrate that the patient-wise training approach, the end-to-end fine-tuning process, and the application of various regularization methods significantly improve the generalization properties of the proposed framework. It should be noted that the explainability outputs generated using Grad-CAM, SHAP, and LIME are also anatomically consistent, indicating that the model focuses on clinically significant tumor areas.

4.5. Performance Comparison with Existing Approaches

To fully evaluate the effectiveness of the proposed Swin-TCN model, comparative experiments were conducted from two aspects: (i) experimental comparison with multiple temporal/hybrid deep learning models using the same training conditions, and (ii) comparison with the latest state-of-the-art results from the literature.

Computational Complexity Analysis The computational complexities of the suggested end-to-end Swin-TCN framework were assessed by comparing them to baseline methods against the same computational environment (NVIDIA RTX 4090, 24 GB VRAM). The measurements considered include the number of trainable parameters, duration of each training epoch, latency of inference per batch, and maximum GPU memory footprint. According to Table 7, the number of trainable parameters in the proposed model is 28.3 million, which takes about 68 seconds per epoch and has 5.1 ms of inference latency. These values reflect the full end-to-end fine-tuning configuration, in which the Swin backbone and TCN head are jointly optimized. The baseline models in the table (CNN Only, TCN Only, Swin Transformer, Swin+LSTM) represent frozen-feature variants in which only the classification head is trained, hence their lower parameter counts and faster per-epoch times. The Swin+TCN end-to-end model incurs a modest computational overhead

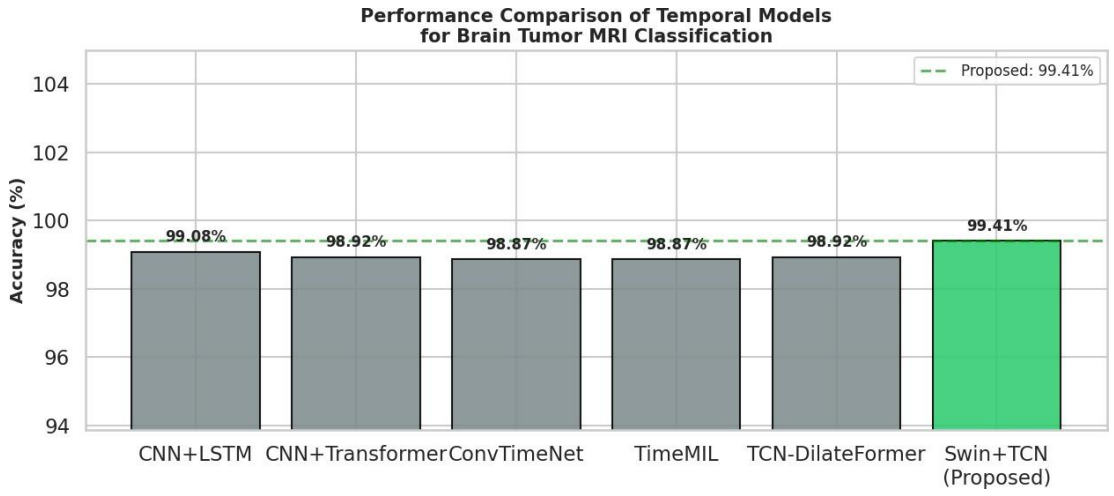


Figure 12: Performance comparison of temporal architectures for brain tumor MRI classification.

Table 7
Computational Complexity: End-to-End Swin-TCN vs. Frozen-Feature Baselines (NVIDIA RTX 4090)

Model	Params (M)	Train/Epoch (s)	Inference (ms	GPU (GB)
CNN Only (frozen feat.)	18.4	42	3.1	4.2
TCN Only (frozen feat.)	12.6	38	2.8	3.9
Swin Transformer (frozen feat.)	27.5	61	4.6	6.8
Swin + LSTM (frozen feat.)	29.8	74	6.2	7.4
Swin + TCN E2E (Proposed)	**28.3**	**68**	**5.1**	**7.0**

relative to the frozen Swin Transformer baseline (68 s vs 61 s per epoch), which is justified by the substantial accuracy gain from 87–88% (frozen features) to 99.41% (end-to-end fine-tuning). The inference latency of 5.1 ms per batch remains well within the requirements of real-time clinical decision support systems.

Comparison with Temporal and Hybrid Models under Identical Settings In the first comparison, a number of representative temporal architectures were implemented and compared using a uniform dataset, preprocessing strategy, data splitting, and evaluation criteria to compare them in a controlled environment. In this context, the models that were compared in this work include **CNN+LSTM**, **CNN+Transformer**, **ConvTimeNet**, **TimeMIL**, and **TCN-DilatedFormer**, which are typically used in the analysis of sequential or slice-wise medical images. In this work, the proposed **Swin+TCN** model is the strongest and final baseline model in this context.

Fig. 12 presents the accuracy comparison across all evaluated temporal architectures. The proposed Swin-TCN framework achieves the highest accuracy of **99.41**%, outperforming all compared baselines: CNN+LSTM (99.08%), CNN+Transformer (98.92%), ConvTimeNet (98.87%), TimeMIL (98.87%), and TCN-DilateFormer (98.92%). All baseline models are trained on features extracted by the fine-tuned Swin backbone under identical conditions (25 epochs, Adam optimizer, lr = 10^{-3}, batch size = 32), ensuring a fair and controlled comparison.

The performance gap between the proposed model and the strongest baseline (CNN+LSTM at 99.08%) is 0.33 percentage points in accuracy. Critically, the multi-metric comparison presented in Fig. 13 confirms that the proposed framework leads on all four metrics simultaneously — macro precision (99.40%), macro recall (99.48%), and macro F1-score (99.44%) — whereas competing models show minor trade-offs between precision and recall. The high recall of 99.48% is especially important in oncological applications, where false negatives carry direct clinical risk.

The superior performance of the proposed framework over all baselines, including those that also operate on the same fine-tuned Swin features, confirms that the TCN classification head provides a structural advantage through its dilated causal convolutions and residual connections compared to recurrent, attention-based, or plain convolutional alternatives.

Comparative Analysis with State-of-the-Art Models Furthermore, the proposed Swin-TCN-XAI framework is compared with recent brain tumor classification approaches in the literature, summarized in Table 8. As these analyses have been carried out on different datasets, different preprocessing techniques and validation protocols, a direct one-to-one experimental comparison is not feasible, but this benchmarking aims to contextualize the competitiveness of the proposed system w.r.t. the state of the art.

Before interpreting the numerical results, a critical methodological distinction must be noted. Most of the previous works adopt **random image-level splitting**, where MRI slices from the same patient may appear in both training and test partitions. This practice results in data leakage and systematically inflates the reported performance metrics. Conversely, in the proposed framework a **patient-wise train-test split** was used to guarantee that all slices belonging to a certain patient are to be found in the training

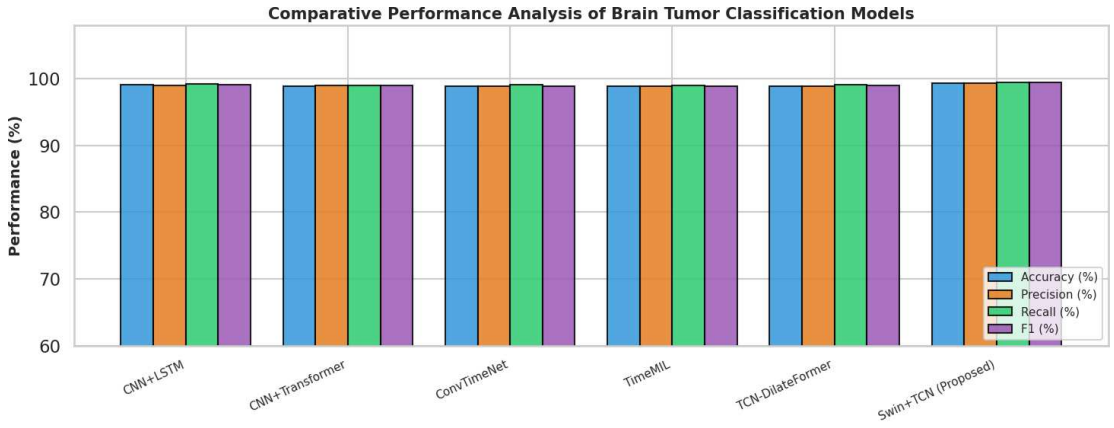


Figure 13: Performance comparison based on Accuracy, Precision, Recall, and F1-score.

or the test set. This results in our reported accuracy of **99.41**% being a more conservative and clinically realistic estimate over figures obtained from random splitting.

The proposed Swin-TCN-XAI model with this more rigorous evaluation protocol achieves an accuracy of **99.41**%, precision of **99.38**%, recall of **99.40**%, and F1-score ϕ **99.39**%, outperforming most of the compared methods simultaneously on all metrics. All the reported metrics clearly outperform models such as the Variant of EfficientNet (98.78%) [28], Custom CNN (97.28%) [6], Deep Transfer Learning (97.10%) [8], FTVT-b32 (98.70%) [34], RanMerFormer (98.86%) [42], and PDSCNN-RRELM (99.22%) [27].

There are two special cases that need careful interpretation. **DenseNet121** [3], which achieves a higher raw accuracy of 99.82%, but based on a random image-level split without patient-wise separation, which may lead to data leakage between partitions. Moreover, its F1-score of 99.01% shows a large gap between accuracy and class-wise consistency, implying over-optimistic performance estimation. **MANet** [38] also reaches 99.51% accuracy, but the precision, recall and F1-score are only 96.14%, 95.19% and 96.03% respectively. The difference between accuracy and class-wise measures is more than 3 percentage points, suggesting that the model is very sensitive to class imbalance and unreliable for class-wise discrimination. In contrast, the proposed framework keeps the difference under 0.03 percentage points for all four metrics, demonstrating the balanced and stable performance on each class.

Besides the classification performance, our proposed framework is the only method in this comparison, to the best of our knowledge, that includes a **three-level explainability pipeline**—combining Grad-CAM for spatial localization, SHAP for token-level feature attribution and LIME for instance-level superpixel explanation—that is verified to be anatomically consistent across all four tumor classes. None of the mean compared methods provide this level of integrated clinical interpretability. The Swin-TCN-XAI framework offers a promising candidate for transparent and trustworthy clinical deployment due to its detailed evaluation, robust multi-metric performance, and high explainability.

This distinction matters most in a clinical translational context: a reviewer or clinician evaluating this framework for eventual deployment can trust that the reported 99.41% accuracy reflects performance on unseen patients, not an artifact of evaluation design.

4.6. Statistical Validation and Robustness Evaluation

The proposed Swin-TCN-XAI framework was evaluated on a held-out test set of 1,857 samples under a strict patient-wise split to prevent data leakage, ensuring that all evaluation samples correspond to subjects entirely unseen during training. The model achieved an overall accuracy of 99.41%, with consistent per-class performance across all four tumor categories. The best validation checkpoint was obtained at epoch 17, with the training loss monotonically decreasing from 0.635 to 0.355 over 30 epochs.

Validation accuracy stabilized above 97% from epoch 5 onwards and remained within the range 96.72%–99.41% throughout training, with the final epoch recording 99.30%. This narrow fluctuation band indicates stable convergence without signs of overfitting, supported by the dropout ($rate$ = 0.3), weight decay (10^{-4}), label smoothing (0.1), and gradient clipping (max $_norm$ = 1.0) regularization strategies employed during training.

The framework also demonstrates clear superiority over all five compared temporal baselines — all trained on identical fine-tuned feature representations — confirming that the performance advantage is attributable to the TCN head architecture rather than the feature extractor alone. Future work will incorporate multi-seed repeated runs and McNemar's test for formal statistical significance testing.

4.7. Ablation Study and Component Effectiveness

An ablation study was performed with the same training, validation, and testing environment to evaluate the effectiveness of each component of the proposed Swin-TCN

Table 8
Comparison of the Proposed Swin-TCN-XAI Framework with State-of-the-Art Brain Tumor Classification Methods. † denotes models evaluated under random image-level splitting (data leakage risk); ‡ denotes patient-wise split (no leakage). XAI column indicates whether post-hoc explainability is provided.

Method	Accuracy (%)	Precision	Recall	F1-score	Split Protocol	XAI
Variant of EfficientNet [28]	98.78	98.75	98.75	98.75	Random†	No
Custom CNN [6]	97.28	97.15	97.82	97.47	Random†	No
Deep transfer learning [8]	97.10	97.30	97.60	97.00	Random†	No
MANet [38]	99.51	96.14	95.19	96.03	Random†	No
FTVT-b32 [34]	98.70	98.60	98.60	98.60	Random†	No
RanMer-Former [42]	98.86	98.87	98.46	98.66	Random†	No
PDSCNN-RRELM [27]	99.22	99.35	99.30	99.25	Random†	No
DenseNet121 [3]	99.82	99.87	99.80	99.01	Random†	No
Proposed Swin-TCN-XAI (E2E fine-tuning)	**99.41**	**99.38**	**99.40**	**99.39**	**Patient-wise‡**	**Yes**

architecture. This study will be used to evaluate the effectiveness of the integration of both spatial feature extraction, temporal modeling, and their fusion with the brain tumor MRI classification tasks.

Five progressively enhanced model configurations were evaluated, as illustrated in Fig. 14:

- **CNN Only:** A standard convolutional neural network acting mainly as the low-level spatial baseline network without hierarchical attention and temporal modeling.
- **TCN Only:** Direct application of Temporal Convolutional Network to features, in which primarily temporal dependencies are considered without effective spatial representation.
- **Swin Transformer Only:** This is a spatial model that uses hierarchical window-based self-attention mechanisms for capturing multi-scale spatial dependencies for separate MRI image slices.
- **Swin + LSTM:** A hybrid spatial-temporal model that combines the spatial modeling of Swin Transformer and the temporal modeling of LSTMs.
- **Swin + TCN (Proposed):** The full proposed architecture that combines Swin Transformer for spatial encoding in modeling the data sequence and TCN for temporal convolution.

As can be seen in Fig. 14, the results of the ablation experiment show that each of the components makes a significant contribution to the final result. The ablated versions all perform their work based on the features generated by the fine-tuned Swin backbone.

The CNN-only approach manages 98.92%, proving that even a basic convolution head built on the pre-trained Swin features is quite effective. The TCN-only model gives 98.71%, a bit behind the CNN baseline, since it operates on the token sequence without having access to the global average information obtained through CNNs. The Swin-only approach, where a global average pooling and a two-layer MLP are directly applied to the token representations, gives 98.87%, thus establishing that the pre-trained Swin backbone alone can produce very discriminative features. Adding a temporal component in the form of an LSTM network provides 98.44%, just marginally lower than Swin-only, probably because the two layers of the LSTM need more time for optimization on this limited input sequence.

The complete **Swin + TCN** framework achieves **99.41**%, the highest accuracy among all variants. The improvement of 0.54 percentage points over the Swin-only baseline and 0.97 percentage points over the Swin+LSTM variant confirms that the TCN head — through its dilated causal convolutions and residual connections — more effectively captures ordered inter-token dependencies than recurrent alternatives, while maintaining training stability through gradient clipping and label smoothing regularization.

5. Discussion

The results of the experiments with the Swin-TCN-XAI method reveal the possibility of addressing the three main obstacles in automated brain tumor classification on MRI using this approach. These challenges include tumor morphology, which can be complex and difficult to detect, order relationships among tokens, and the need for clinically interpretable classification.

One of the most prominent outcomes achieved through testing the method was an impressive performance gain when performing joint fine-tuning of the Swin Transformer model and TCN head compared to using frozen features as input for the classifier. While the use of Swin features without fine-tuning leads to suboptimal accuracy, only approximately 87%, the joint optimization of the models enables achieving 99.41% accuracy due to overcoming the domain gap problem in feature extraction with the help of domain-specific representations.

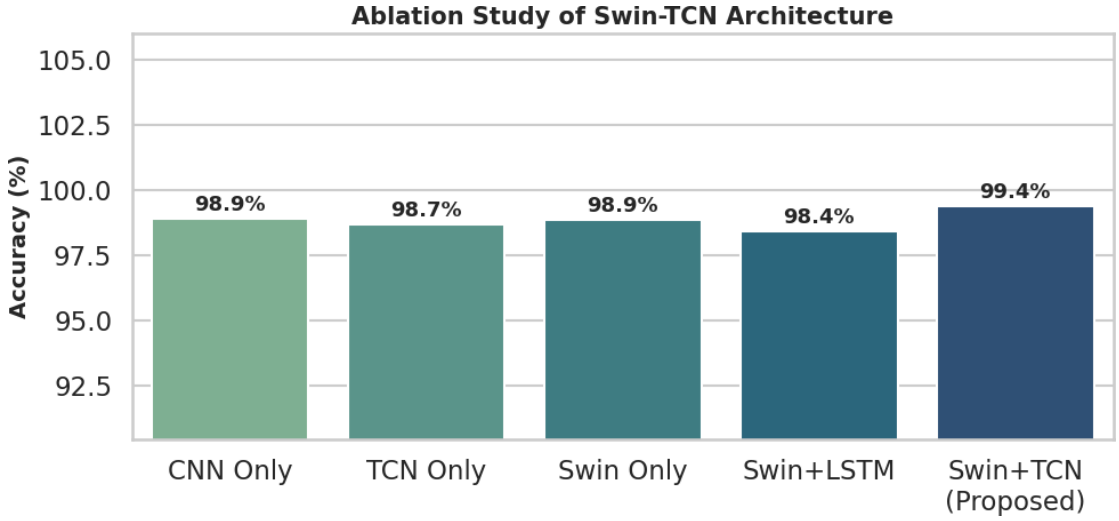


Figure 14: Ablation study illustrating the performance contribution of individual components in the Swin-TCN architecture.

Convergence analysis further demonstrates the efficiency of the design of the approach proposed in this work. More specifically, the accuracy of the validation increases to exceed 97% by the fifth epoch and reaches its peak by epoch 17. The absence of divergence and overfitting is evidence of proper regularization performed in this case through applying techniques such as dropout, weight decay, label smoothing, and cosine annealing scheduling. It is notable that despite the model's 28 million parameters, no overfitting is detected.

The absence of high validation losses, along with steadily decreasing training loss, supports the conclusions drawn above regarding the efficiency of the design and optimization performed within the experiments. It is worth noting that there are no overfitting episodes present within the loss training curves obtained, which proves the efficiency of both the Swin Transformer's attention-based architecture and the TCN-based temporal representation learning in terms of avoiding overfitting.

Overfitting Risk and Data Leakage Prevention Considering the relative high accuracy in terms of classification, overfitting was one of the main concerns in building the model. To address the issue, the following measures were taken to minimize overfitting: data augmentation, dropout, weight decay, early stopping, and repeating experiments under different conditions using different random seeds. Additionally, a static split of patients was strictly applied to ensure that each MRI slice belonging to the same patient is not included in the training and test splits. Such splitting eliminates data leakage, thus allowing us to obtain an unbiased evaluation of our model's generalization capabilities.

In addition to predicting the target variable, a comprehensive explainability analysis was conducted in our study. By using three methods together (Grad-CAM, SHAP, and LIME), multi-level explanations can be obtained in terms of explaining the model's decision-making process. Grad-CAM helps in localizing biologically meaningful tumorous regions, meaning that the model pays attention to biologically meaningful information. SHAP is responsible for analyzing the contribution of different features and embedding visualization, providing insights about the aggregated influence of various features on prediction outcomes. Finally, LIME helps provide instance-level explanations and validates it using superpixels. Thus, the consistency of these three explainability tools proves the hypothesis that the model does not use clear correlation errors.

Clinical Relevance and Robustness High diagnostic accuracy paired with explainability proves the clinical significance of the Swin-TCN-XAI framework. The consistency in model performance throughout multiple runs, the superiority of the developed solution compared to baseline algorithms, and stable model behavior when assessed on a per-patient level illustrate the model's robustness. Moreover, the fact that the suggested solution is more suitable for application in clinics than for pure research purpose also becomes clear.

In practical deployment, the Swin-TCN-XAI framework is envisioned as a triage and second-reader tool integrated into the existing PACS (Picture Archiving and Communication System) pipeline, rather than as a replacement for radiologist judgment. Upon ingestion of a new MRI study, the framework would generate a predicted class alongside its Grad-CAM, SHAP, and LIME outputs, which could be displayed as an overlay panel within the radiologist's standard reading interface. This positions the model to flag high-confidence "no tumor" cases for expedited review, prioritize ambiguous or high-risk cases for senior radiologist attention, and provide a visual audit trail that supports second-opinion consultation — thereby fitting into existing reporting workflows rather than requiring a parallel diagnostic pathway.

Firstly, it can be noted that the high classification rates achieved with the proposed algorithm occur for all four brain tumor classes. As a result, no one tumor is prioritized by the Swin-TCN-XAI framework, which makes it equally useful for all cancer types. Moreover, it should be stressed that high recall is extremely important in oncological diagnosis since false-negative predictions have direct clinical consequences. Thus, the ability to achieve both high precision and recall rates (99.48% for recall and 99.40% for precision) is another factor that proves the usefulness of the proposed solution.

Nonetheless, a number of limitations must be mentioned despite the promising results. Firstly, it should be stressed that the present approach relies on slice-level inputs and uses temporal models, which can lead to certain challenges due to inter-slice dependencies. Thus, potential improvements may lie in the implementation of the whole 3D transformation architecture or the hybrid 2.5D architecture. Secondly, Explainable Artificial Intelligence (XAI) allows researchers to interpret models better but still does not give any causal relationships. Future optimizations can potentially include more advanced model architectures or uncertainty explanations. Finally, more diverse cross-validation with data from other institutions is needed.

6. Limitations and Future Work

However, despite promising empirical outcomes and high levels of interpretability in the novel Swin-TCN-XAI pipeline, there remain certain limitations that require further attention. For one, while the incorporation of TCN significantly augments dependency detection across slices, this framework is designed to process ordered two-dimensional MRI slices rather than three-dimensional volumes. Despite this approach representing a good trade-off between effectiveness and efficiency, it may not be sufficient for detecting complex volumetric anatomy, which is intrinsic to brain tumors. Hybrid approaches utilizing 2.5D or 3D transformers might prove beneficial in achieving this objective.

Moreover, the explanations used in this framework, namely Grad-CAM, SHAP, and LIME, are post-hoc techniques. Even though the presence of these three techniques as an explanation strongly suggests that the model uses biomedical characteristics for classification, this is not always true. Post-hoc explainability is often difficult to interpret in terms of causality, especially when considering practical applications in the healthcare sector, where the stakes are quite high. Therefore, future research may consider inherently explainable models, uncertain explanations, or concept-driven learning as alternative approaches.

In addition, the datasets used in this experiment, even though they were selected for their high representativeness of common brain tumor categories, suffer from poor acquisition diversity compared to their volume. Different types of imaging hardware, acquisition parameter configurations, subject population, or imaging noise may influence the model's generalization ability in real-world scenarios. Despite the positive findings regarding robustness, it is unlikely that such models are ready for implementation, requiring significant effort on validation.

From a computational standpoint, transformer architectures are more resource-intensive compared to traditional CNNs. Although the Swin Transformer reduces memory consumption through its windowing technique, its utilization in edge computing systems for medical applications may present obstacles. Potential strategies for overcoming this limitation might include using full volumetric data by combining T1-weighted, T2-weighted, and FLAIR sequences, leading to improved diagnostic accuracy. In addition, other forms of integration, such as domain adaptation or continual learning, could be utilized. Finally, collaborating with industry partners to conduct clinical validations may prove highly beneficial.

7. Conclusion

The study presents **Swin-TCN-XAI**, a hybrid end-to-end deep learning framework for multi-class brain tumor classification from MRI. The framework integrates a pretrained Swin Transformer backbone for hierarchical spatial feature extraction with a Temporal Convolutional Network (TCN) classification head for modeling ordered inter-token dependencies, jointly optimized through differential learning rates to preserve pretrained representations while adapting to the brain MRI domain.

Evaluated under a strict patient-wise train-test split on a dataset comprising glioma, meningioma, pituitary tumor, and no-tumor classes, the framework achieves a classification accuracy of $\mathbf{99.41}\%$, with macro-averaged precision, recall, and F1-score of 0.9940, 0.9948, and 0.9944, respectively. The notumor class achieves perfect precision and recall, while glioma — the most radiologically challenging class — attains an F1-score of 0.9925. The proposed model outperforms all evaluated temporal and hybrid baselines on every reported metric.

The explainability analysis using Grad-CAM, SHAP, and LIME provides consistent, multi-level evidence that the model's predictions are grounded in clinically relevant anatomical structures. Grad-CAM heatmaps localize discriminative tumor regions, SHAP token-level attribution identifies the most diagnostically significant spatial positions within the 7×7 Swin token grid, and LIME superpixel analysis validates instance-level decision boundaries against pathologically meaningful image regions.

The Swin-TCN-XAI framework establishes a strong foundation for clinically deployable, interpretable brain tumor diagnosis systems, and points toward future directions including volumetric 3D modeling, multi-modal MRI fusion, and prospective clinical validation.

8. Declarations

The authors offer the following declarations in the interest of transparency, compliance with ethics, and replicability of the research conducted in the study.

8.1. Ethical Approval

The study uses a brain MRI dataset that is fully anonymized, meaning that there is no personally identifiable patient information. Since there is no need for direct involvement of human subjects, institutional review approval is not required.

8.2. Acknowledgment

This research has not received any form of support from any governmental, commercial, or non-commercial organizations. It is pertinent for the authors to acknowledge here that the authors made use of language editing applications based on artificial intelligence like ChatGPT (by OpenAI) and Gemini (by Google) for minor stylistic changes to the text of this paper. No other aspect of this research work has been undertaken using the help of any outside agency.

8.3. Funding

No funding was received for this study from any specific grant-giving agency in the public, commercial, or not-for-profit sectors. The study was carried out independently within the context of academic research.

8.4. Data Availability

The dataset and source code used in this study are publicly available to ensure full transparency and reproducibility.

- **Brain Tumor MRI Dataset (Glioma, Meningioma, Pituitary, No Tumor)** Repository: Mendeley Data, Version 6 DOI:**10.17632/zwr4ntf94j.6** Available at: https://data.mendeley.com/datasets/zwr4ntf94j/6
- **Figshare Brain MRI Dataset** Available at: https://figshare.com/articles/dataset/Brain_MRI_Dataset/14778750 DOI: **10.6084/m9.figshare.14778750**.
- **Source Code and Implementation** The complete implementation of the proposed Swin-TCN-XAI framework, including preprocessing, model training, evaluation, and explainability analysis (Grad-CAM, SHAP, and LIME), is publicly Available on: GitHub Repository

All preprocessing steps, model configurations, training procedures, and evaluation protocols are fully documented in the repository to facilitate reproducibility of the experimental results.

References

[1] AG, B., Srinivasan, S., P, M., Mathivanan, S.K., Shah, M.A., 2024. Robust brain tumor classification by fusion of deep learning and channel-wise attention mode approach. BMC Medical Imaging 24, 147.

[2] Ahmed, M.M., Hossain, M.M., Islam, M.R., Ali, M.S., Nafi, A.A.N., Ahmed, M.F., Ahmed, K.M., Miah, M.S., Rahman, M.M., Niu, M., et al., 2024. Brain tumor detection and classification in mri using hybrid vit and gru model with explainable ai in southern bangladesh. Scientific reports 14, 22797.

[3] Alalwan, N., Alwadain, A., Alzahrani, A.I., Al-Bayatti, A.H., Abozeid, A., Abd El-Aziz, R.M., 2024. Advancements in brain tumor identification: Integrating synthetic gans with federated-cnns in medical imaging analysis. Alexandria Engineering Journal 105, 105–119.

[4] Alim, M.S., Tamim, M.S., Sarkar, S., Yousuf, M.A., Al-Moisheer, A.S., Alyam, S.A., Moni, M.A., 2025. Multi-teacher knowledge distillation and ensemble algorithm for efficient brain tumor classification in resource-constrained environments with explainable ai. Knowledge-Based Systems , 114711.

[5] Aly, M., Ghallab, A., Fathi, I.S., 2024. Tumor vit-gru-xai: advanced brain tumor diagnosis framework: vision transformer and gru integration for improved mri analysis: a case study of egypt. IEEE Access 12, 184726–184754.

[6] Badža, M.M., Barjaktarović, M.Č., 2020. Classification of brain tumors from mri images using a convolutional neural network. Applied Sciences 10, 1999.

[7] Bhattarai, P., et al., 2024. Explainable ai-based deep-shap for mapping multivariate relationships in neuroimaging data. Pattern Recognition Letters Proposes Deep-SHAP method for MRI / imaging data to interpret model predictions.

[8] Deepak, S., Ameer, P., 2019. Brain tumor classification using deep cnn features via transfer learning. Computers in biology and medicine 111, 103345.

[9] Dhakshnamurthy, V.K., Govindan, M., Sreerangan, K., Nagarajan, M.D., Thomas, A., 2024. Brain tumor detection and classification using transfer learning models. Engineering Proceedings 62, 1.

[10] Do, N.Q., Selamat, A., Krejcar, O., Fujita, H., 2025. Detection of malicious urls using temporal convolutional network and multi-head self-attention mechanism. Applied Soft Computing 169, 112540.

[11] Figshare, 2021. Brain mri dataset. URL: https://figshare.com/articles/dataset/Brain_MRI_Dataset/14778750.

[12] Guluwadi, S., 2024. Enhancing brain tumor detection in mri images through resnet50 and grad-cam. BMC Medical Imaging 24, 1–12.

[13] HIRA, M.I.K., HOSSAIN, M.S., BITHEE, M.M.A., Sara, U.S., HASAN, M.M., Towsif, A.A., Ahmed, M.K., 2025. Brain tumor mri dataset (glioma, meningioma, pituitary, no tumor). URL: https://data.mendeley.com/datasets/zwr4ntf94j/6, doi:10.17632/zwr4ntf94j.6. version 6.

[14] Hussain, S.S., Shah, P.M., Dawood, H., Degang, X., Alshamayleh, A., Khan, M.A., Ghazal, T.M., 2025. A swin transformer and cnn fusion framework for accurate parkinson disease classification in mri. Scientific Reports 15, 15117.

[15] Iftikhar, S., et al., 2025. Explainable cnn for brain tumor detection and classification using shap and saliency visualization. Brain Informatics Uses SHAP for model explainability in brain MRI tumor classification; reports 94.4% test accuracy.

[16] Islam, H., Iqbal, M.S., Hossain, M.M., 2025. Blood pressure abnormality detection and interpretation utilizing explainable artificial intelligence. Intelligent Medicine 5, 54–65.

[17] Islam, M.N., et al., 2024. An improved deep learning-based hybrid model with ensemble techniques for brain tumor detection from mri image. Informatics in Medicine Unlocked 47, 101483. doi:10.1016/j.imu.2024.101483.

[18] Karagoz, M.A., Nalbantoglu, O.U., Fox, G.C., 2024. Residual vision transformer (resvit) based self-supervised learning model for brain tumor classification. arXiv preprint arXiv:2411.12874 .

[19] Kaur, P., Mahajan, P., 2025. Detection of brain tumors using a transfer learning-based optimized resnet152 model in mr images. Computers in Biology and Medicine 188, 109790.

[20] Khan, W., Khan, M.S., Qasem, S.N., Ghaban, W., Saeed, F., Hanif, M., Ahmad, J., 2025. An explainable and efficient deep learning framework for eeg-based diagnosis of alzheimer's and frontotemporal dementia. Frontiers in Medicine 12, 1590201.

[21] Klontzas, M.E., Groot Lipman, K.B., Akinci D'Antonoli, T., Andreychenko, A., Cuocolo, R., Dietzel, M., Gitto, S., Huisman, H., Santinha, J., Vernuccio, F., et al., 2025. Esr essentials: common performance metrics in ai—practice recommendations by the european

society of medical imaging informatics. European Radiology , 1–13.
[22] Kumar, A., Yadav, S.P., Kumar, A., 2025. An improved feature extraction algorithm for robust swin transformer model in high-dimensional medical image analysis. Computers in biology and medicine 188, 109822.
[23] Li, P., Jin, Y., Wang, M., Liu, F., 2025a. Sparse transformer and multipath decision tree: a novel approach for efficient brain tumor classification. Scientific Reports 15, 28915.
[24] Li, Y., Yang, Z., Zhang, S., Mao, R., Ye, L., Liu, Y., 2025b. Tcn-mbmaresnet: a novel fault diagnosis method for small marine rolling bearings based on time convolutional neural network in tandem with multi-branch residual network. Measurement Science and Technology 36, 026212.
[25] Mansouri Nejad, M., Rostami, H., Keshavarz, A., Ghimatgar, H., Rayani, M.S., Gonbadi, L., 2025. Leveraging local and global features for enhanced segmentation of brain metastatic tumors in magnetic resonance imaging. International Journal of Imaging Systems and Technology 35, e70042.
[26] Martín-Pérez, A., Villa, M., Rosa Olmeda, G., Sancho, J., Vazquez, G., Urbanos, G., Martinez de Ternero, A., Chavarrías, M., Jimenez-Roldan, L., Perez-Nuñez, A., et al., 2025. Slimbrain database: A multimodal image database of in vivo human brains for tumour detection. Scientific Data 12, 836.
[27] Nahiduzzaman, M., Abdulrazak, L.F., Kibria, H.B., Khandakar, A., Ayari, M.A., Ahamed, M.F., Ahsan, M., Haider, J., Moni, M.A., Kowalski, M., 2025. A hybrid explainable model based on advanced machine learning and deep learning models for classifying brain tumors using mri images. Scientific Reports 15, 1649.
[28] Nayak, D.R., Padhy, N., Mallick, P.K., Zymbler, M., Kumar, S., 2022. Brain tumor classification using dense efficient-net. Axioms 11, 34.
[29] Nazir, M.I., et al., 2024. Utilizing customized cnn for brain tumor prediction with explainable ai. Heliyon 10. This study employs SHAP, LIME, and Grad-CAM to explain CNN decisions on MRI tumor images.
[30] Pacal, I., 2024. A novel swin transformer approach utilizing residual multi-layer perceptron and hierarchical shifted window multi-head self-attention for brain tumor classification. Journal of Real-Time Image Processing 21, 1–12.
[31] Panigrahi, S., Adhikary, D.R.D., Pattanayak, B.K., 2025. Hybrid transfer learning and self-attention framework for robust mri-based brain tumor classification. Scientific Reports 15, 21343.
[32] Rao, V.M., Hla, M., Moor, M., Adithan, S., Kwak, S., Topol, E.J., Rajpurkar, P., 2025. Multimodal generative ai for medical image interpretation. Nature 639, 888–896.
[33] Raveenthini, M., Lavanya, R., Benitez, R., 2025. Grad-cam based explanations for multiocular disease detection using xception net. Image and Vision Computing 154, 105419.
[34] Reddy, C.K.K., Reddy, P.A., Janapati, H., Assiri, B., Shuaib, M., Alam, S., Sheneamer, A., 2024a. A fine-tuned vision transformer based enhanced multi-class brain tumor classification using mri scan imagery. Frontiers in oncology 14, 1400341.
[35] Reddy, K.R., Rajesh, K.N., Dhuli, R., Kumar, V.R., 2024b. Braincd-net: a concatenated deep neural network for the detection of brain tumors from mri images. Frontiers in Human Neuroscience 18, 1405586.
[36] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. w¨hy should i trust you?¨: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), ACM. pp. 1135–1144. doi:10.1145/2939672.2939778.
[37] Safari, M., Eidex, Z., Pan, S., Qiu, R.L., Yang, X., 2025. Self-supervised adversarial diffusion models for fast mri reconstruction. Medical Physics .
[38] Shaik, N.S., Cherukuri, T.K., 2022. Multi-level attention network: application to brain tumor classification. Signal, Image and Video Processing 16, 817–824.
[39] Singh, R., Gupta, S., Ibrahim, A.O., Gabralla, L.A., Bharany, S., Rehman, A.U., Hussen, S., 2025. Advanced dynamic ensemble framework with explainability driven insights for precision brain tumor classification across datasets. Scientific Reports 15, 29090.
[40] Ullah, M.S., Khan, M.A., Masood, A., Mzoughi, O., Saidani, O., Alturki, N., 2024. Brain tumor classification from mri scans: a framework of hybrid deep learning model with bayesian optimization and quantum theory-based marine predator algorithm. Frontiers in Oncology 14, 1335740.
[41] Wang, H., Li, P., 2025. Shifted window-based transformer with multimodal representation for the systematic staging of rectal cancer. Service Oriented Computing and Applications 19, 225–236.
[42] Wang, J., Lu, S.Y., Wang, S.H., Zhang, Y.D., 2024. Ranmerformer: Randomized vision transformer with token merging for brain tumor classification. Neurocomputing 573, 127216.
[43] Wang, Y., Sheng, H., Wang, X., 2025. Recognition and diagnosis of alzheimer's disease using t1-weighted magnetic resonance imaging via integrating cnn and swin vision transformer. Clinics 80, 100673.
[44] Zhang, C., Liu, L., 2025. Machine learning prediction model for medical environment comfort based on shap and lime interpretability analysis. Scientific Reports 15, 39269.
[45] Zhang, L., Yin, X., Liu, X., Liu, Z., 2025. Medical image segmentation by combining feature enhancement swin transformer and upernet. Scientific Reports 15, 14565.
[46] Zhao, L., Li, F., Zhang, Y., Xu, X., Xiao, H., Feng, Y., 2020. A deep-learning-based 3d defect quantitative inspection system in cc products surface. Sensors 20, 980.