Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 3–12
DOI: 10.15439/2025F0901
ISSN 2300-5963 ACSIS, Vol. 43

Multiscale MoE: A Mixture of Experts Framework with Attention-Driven Multi-Scale Learning for Brain Tumor Classification

Ameer Hamza, Robertas Damaševičius
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania Email: ameer.hamza@ktu.edu, robertas.damasevicius@ktu.lt

***Abstract*—The impact of brain tumors as a global health concern is due to their aggressive behavior, high mortality, and complexities in their diagnosis. While MRI remains the gold standard for identifying, monitoring, and detecting brain tumors, automated classification methods encounter many complications with respect to the diverse morphologies of tumors, similarities in their imaging features, and the potential variability in imaging conditions.CNNs can capture spatial hierarchies, but cannot generalize effectively and ViTs rely on the context to characterize the image modalities which means that, whilst they address some deficiencies of CNNs, they require extensive data and computational resources. To remedy some of the issues that each approach presents, we present multiscale MoE that leverages CNNs and attention-oriented modules. The proposed architecture uses multi-scale feature extraction, channel-spatial attention, and dynamic expert routing, which adequately collects tumor-specific features efficiently. We applied two different publicly available datasets, namely the Bangladesh Brain Cancer MRI and Figshare Brain Tumor dataset. For the Bangladesh dataset, the proposed model achieved overall accuracy of 96.92% and for FigShare dataset, the highest results achieved 96.42% accuracy. In contrast to state-of-the-art models, multiscale MoE achieved the highest testing accuracy 96.14%, and the lowest Brier score 0.0603. The proposed model has shown to have balanced classification results across the tumor classes and reduced the number of false predictions whilst maintaining efficient computational performance and thus has the potential to provide a valuable resource for clinical practice with respect to real-time applications.**

I. Introduction

BRAIN tumors, whether they are primary or metastatic, are a major global health challenge because of their high mortality rates and difficult clinical presentations [1]. Symptoms of brain tumors vary broadly sometimes dependent on location, size, or growth speed and can include debilitating headaches, seizures, neurological deficits, cognitive deficits, and problems with speech or vision [2]. Global health data suggests that this burden is increasing: in 2020, an estimated 308,000 new brain and CNS tumors were diagnosed around the world which led to approximately 251,000 deaths [3]. In the United States, the estimated number of new malignant brain tumors for 2023 to 2025 is expected to reach nearly 24,000 new cases per year with the projection of approximately 18,330 brain tumor related deaths in 2025. Despite advances in their treatment, the overall five-year survival rate is still below 35%, with even lower rates for more aggressive brain tumor types like glioblastoma at close to 7%. This clearly highlights the need for better methods for diagnosis and treatment [4].

Brain tumors are usually approached with a multi-modality approach involving surgical excision, chemotherapeutics, and radiotherapy, which is additionally accompanied with corticosteroids to control intracranial pressure, and anticonvulsant medications to control seizures [5]. Currently, MRI imaging is the gold standard in brain tumor imaging, used for identifying, localizing, and monitoring tumors because it is non-invasive and has greater spatial resolution than the other imaging modalities, and it gives the surgeon an idea of structural and functional information regarding tumor heterogeneity [6]. The advent of automated imaging approaches, driven by MRI precision, has prompted an interest in a collection of imaging variants to assist radiologists in classifying brain tumors based on multimodal MR imaging datasets and to provide images a radiologist would have a unique understanding of and the stages could potentially be determined as well [7].

AI and deep learning specifically, has arisen as a powerful remedy to this issue with unique advancements in assessing the vast amounts of medical images we produce each day [8]. CNNs are the current state-of-the-art for classifying brain tumors as they are able to learn spatial hierarchies and differences based on the spatial relationships between features in MRIs [9]. Recently, ViTs have emerged as a promising rival by modelling global dependencies and self-attention across patch data. Regardless, both approaches address limitations [10]. CNNs rely on intensive training with finite and homogeneous training sets and their performance is symptomatic of their training process, as their ability to classify imaged tumors are less generalizable across imaging conditions[11]. ViTs struggle as they require large amounts of training data and rigorous computational resources to offset their limitations. They also depend upon well-performing patch models over large regions of space and subject to varying spatial biases, which is more frequently defined differently in medical domains because of variability and size [12].

Another problem is the need to develop explainable AI models, which could provide the explanation and justification of its decision [13], [14].

In order to overcome these challenges and limitations, the Mixture of Experts (MoE) framework serves as a promising

path forward because it combines multiple sub-model, allowing each expert to focus on different features of the input such as shape of tumor, texture, and surrounding tissue context. In MoE architecture, each expert focuses on a different feature representation, while a gating mechanism selects the best expert for a given input and its properties. MoE can even manage inter-patient differences, irregular morphologies of tumors, and inconsistencies in medical imaging that often complicate single model approaches. By integrating the advantages of both CNNs and transformers in to a modular framework, the MoE framework can improve classification and generalizability and can improve interpretability of decision making processes. Thus, the Mixture of Experts approach not only addresses the various issues regarding current approaches to automated brain tumor classification but also opens up the possibility for more reliable, rigorously, sensitive and clinically useful AI-supported diagnostic tools. In this work, we proposed a multiscale MoE for the precise classification of brain tumors.

The contributions of this study are as follows:

- A novel Multiscale MoE architecture consist of five experts, multiscale feature extraction, channel-spatial attention, and dynamic expert routing mechanism has been proposed.
- Comprehensive evaluation has been conduction based on two benchmark MRI datasets based on calibration, and inference speed comapred to the SOTA methods.
- Expert utilization analysis revealing dataset-specific specialization patterns, improving interpretability of decision-making.

II. Related Works

In recent years, a number of researchers have proposed several methods for using MRI scans to detect brain cancers [15]. These methods include both conventional machine learning algorithms and deep learning models [16]. Here is a presentation of the pertinent research on brain tumor identification utilizing brain tumor datasets [17]. For example Muhammad et al.[18] presented a comprehensive examination of the various grades used to classify brain tumors. A detailed explanation was given of the procedures used to categorize brain tumors (BTCs), including preprocessing the tumor, determining deep learning features, and classification. They discussed the particular limitations and achievements of the existing deep learning techniques for Bitcoin. The importance of transfer learning for deep learning feature extraction was also covered.

Narmatha et al. [19] presented a methods for classification and segmentation using a fuzzy brain-storm optimization algorithm. With this approach, the target brain cluster's greatest priority is provided by the storm optimization. To find the best answer, the fuzzy procedure is iterated several times. An accuracy of 93.85% is given for the experimental procedure, which was conducted using the BRATS2018 dataset.

Sajad et al.[20] presented a CNN-based multimodal tumor categorization method. They first separated the tumor regions in the MRI data using CNN. They then performed an extensive data augmentation to train an efficient CNN model. Later, they improved the CNN model that was already trained using improved brain data. Tumors were classified using the final layer of the presented method, and it was shown that improved data yielded superior results on the selected datasets.

Mzoughi et al. [21] presented a method for making neuroradiology simple. This study's primary goal was to use volumetric 3D MRI to detect brain tumors. The authors classified the tumors using a Multiscale 3D CNN architecture to increase process efficiency. By using tiny kernels, the suggested approach can lessen the weight of both local and global information. Additionally, the data augmentation technique is used to improve model training. Ultimately, they used experimental findings to demonstrate the effects of data augmentation.

III. Proposed MultiScale MoE

Convolutional Neural Networks have been used for image classification tasks for more than a decade now, demonstrating great potential in various domains particularly in medical imaging. Despite their good performance, there were some issues in traditional CNNs which needed to be addressed such as their inability to extract global patterns, handle diverse features and focus on important regions. Although these issues were addressed individually through attention modules and multi-scale architectures in different studies but there is no comprehensive model that can tackle all the issues together. In this paper, we proposed a new model which can handle all these limitations single handedly by incorporating attention modules, multi scale feature extraction block and Multi scale Mixture of experts in a single architecture to ensure efficient and robust image classification for diverse feature maps.

A detailed diagram of proposed architecture is shown in Figure 1.

Stage 1 begins with initial feature extraction using a 7×7 convolution, batch normalization, ReLU activation, and max pooling, followed by the first MoE block consisting of five parallel expert networks and a dynamic router that assigns adaptive weights to each expert's output. The aggregated expert features are refined through a Channel Attention module, which applies both max-pooling and average-pooling operations to emphasize important channels, and a Spatial Attention module, which uses mean and max pooling across spatial dimensions to highlight critical spatial regions. An Auxiliary Classifier is connected at the end of this stage to provide intermediate supervision.

The proposed model starts with an initial feature extraction block which lays the foundation for deep feature extraction. This block accepts an input of size $(224 \times 224 \times 3)$ and passes it to a 7×7 convolutional layer with a stride of 2 to process low-level features such as textures and edges. It is followed by a batch normalization layer β to normalize the features for stabilized training and a ReLU activation function σ to introduce non-linearity in order to capture complex features. After that, a max pooling layer M_p with a kernel size of 3×3 and a stride of 2 is applied to reduce the spatial dimensions

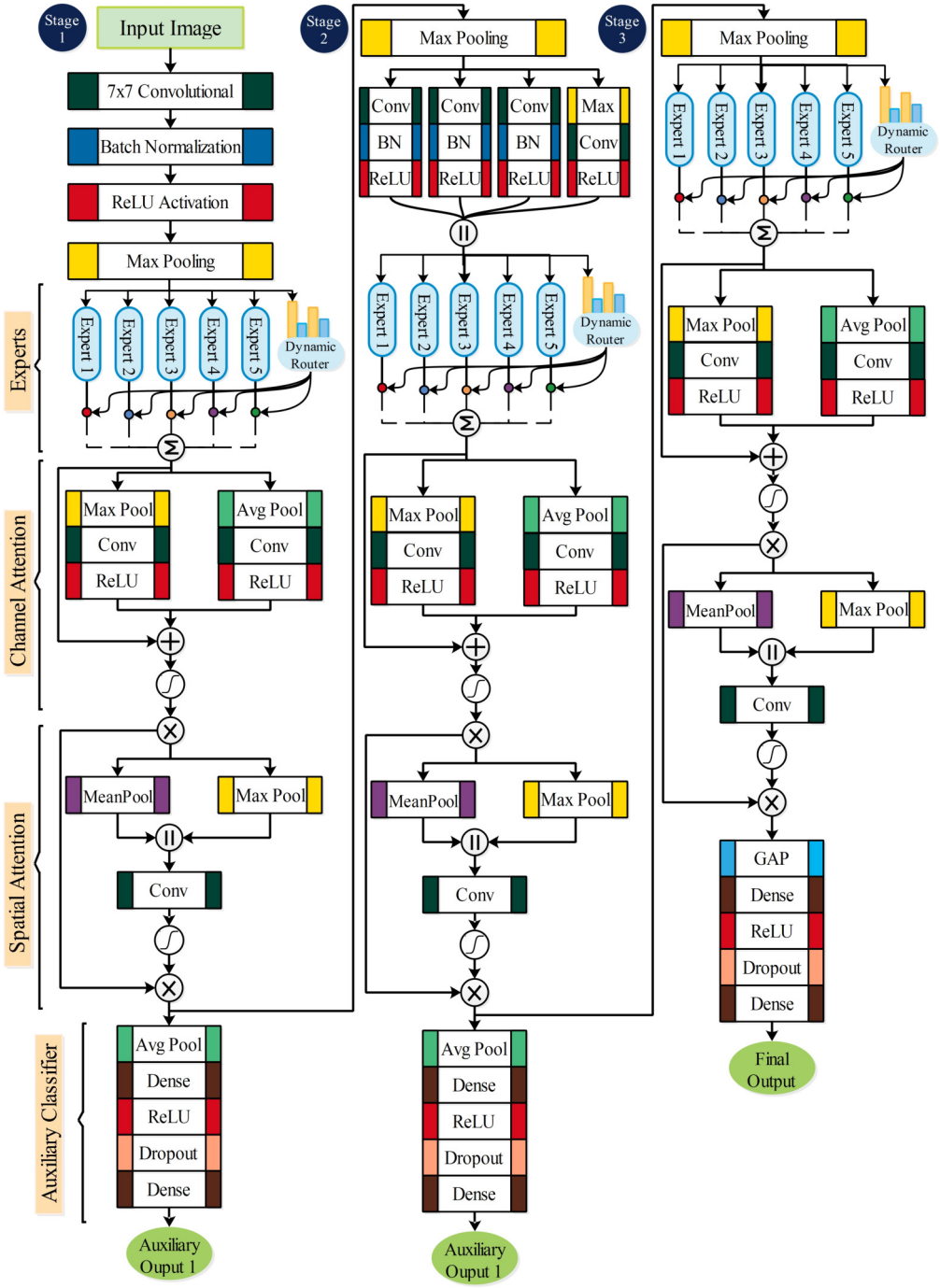

Fig. 1. Detailed architecture of the proposed Multiscale Mixture of Experts (MoE) framework for brain tumor classification. The model operates in three sequential stages, each comprising specialized processing modules.

from 224×224 to 56×56. However, after this block, the number of channels is increased from 3 to 64 for deep feature processing. Mathematically, this block can be represented as:

$$Z_1 = M_p\left(\sigma\left(\beta\left(W_1 * X\right) + b_1\right)\right) \tag{1}$$

Here, X represents the input tensor, $*$ represents the convolutional operation, W_1 represents the convolutional weights, and b_1 represents the bias term.

Stage 2 introduces a multiscale feature extraction block with parallel convolutional layers of varying kernel sizes and pooling operations to capture features at different spatial resolutions. The output is processed by a second MoE block with its dynamic router, followed again by channel and spatial

attention modules and a second auxiliary classifier.

Mixture of Experts: After initial feature extraction, the feature map is passed to the first block of the Mixture of Experts to extract complex features. This block is composed of 5 expert blocks and a dynamic router which assigns weights to each expert block. An expert block is a convolutional neural network that contains several convolutional, batch normalization, and activation layers to extract specialized features based on the input image.

In an expert block, the feature map is first passed through two 3×3 convolutional layers to extract features, where each convolutional layer is followed by a batch normalization layer to normalize the activations. A ReLU activation function is applied at the end to introduce non-linearity. All the expert blocks share the same architecture but have different weights that were initially assigned during training, making them specialized for different types of tumors. An expert block can be defined as:

$$E_i = \sigma\left(\beta\left(W^{''} * \beta\left(W^{'} * Z_1 + b^{'}\right) + b^{''}\right)\right) \tag{2}$$

Here, $W^{'}$ and $W^{''}$ represent the weights of both convolutional layers, and E_i represents the output of the i-th expert block.

On the other hand, a router is composed of an adaptive average pooling layer to summarize spatial information and two dense layers that produce a probability distribution over the 5 expert blocks. A ReLU activation is also applied between both dense layers. It can be defined as:

$$R_i = \psi\left(W_2 \cdot \sigma\left(W_1 \cdot A_p(Z_1)\right)\right) \tag{3}$$

Here, A_p represents adaptive average pooling, W_1 and W_2 represent the weights of the dense layers, and ψ represents the SoftMax activation function which converts the router weights into a probability distribution.

In the MoE block, the input is passed to each expert block and the dynamic router at the same time. The router analyzes the image and assigns weights to each expert block according to their specificity, ensuring that the most relevant experts receive the maximum weightage. The input is processed by each expert individually, and their outputs are multiplied with their respective weights and then added to generate the final feature representation. It can be defined as:

$$\hat{Y} = \sum_{i=1}^{5} R_i \cdot E_i(Z_1) \tag{4}$$

The block diagrams of the expert block and dynamic router are shown in Figure 2 and Figure 3, respectively.

Stage 3 incorporates deeper feature representations via a third MoE block, followed by attention modules and a Final Classifier consisting of global average pooling (GAP), fully connected layers, ReLU activation, dropout regularization, and a final dense layer for tumor type prediction. The architecture's modular design allows dynamic expert utilization, multi-scale feature capture, and attention-guided refinement, enabling robust and efficient classification across diverse MRI tumor appearances.

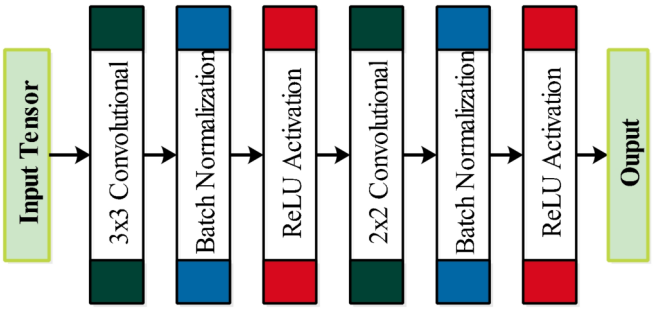

Fig. 2. Structure of an individual expert block used in the proposed Multiscale Mixture of Experts (MoE) framework. The block receives the input tensor and processes it through a sequence of layers: a 3×3 convolutional layer for local feature extraction, followed by batch normalization to stabilize training and ReLU activation to introduce non-linearity. A second convolutional layer with a 2×2 kernel is then applied for further feature refinement, again followed by batch normalization and ReLU activation. The output of this block provides specialized feature representations that contribute to the MoE's adaptive expert routing mechanism.

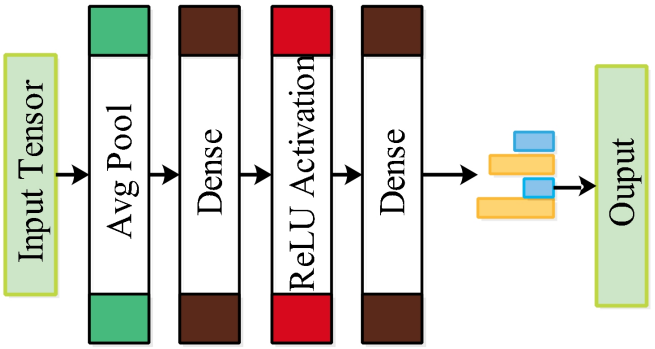

Fig. 3. Architecture of the dynamic router module in the proposed Multiscale Mixture of Experts (MoE) framework. The router receives the input tensor and applies an average pooling operation to capture global spatial information. The pooled features are passed through a fully connected (dense) layer, followed by a ReLU activation to introduce non-linearity, and then through a second dense layer to generate expert routing scores. These scores are normalized into a probability distribution, which determines the relative contribution of each expert in the MoE to the final feature aggregation.

Channel and Spatial Attention: The output of the Mixture of Experts block is passed to the channel attention module to calculate the channel-wise importance and emphasize the important features. For this purpose, the feature map is passed through two parallel blocks simultaneously. Each block is composed of a pooling layer, two convolutional layers, and a ReLU activation function. The only difference is that one block contains an average pooling layer while the other contains a max pooling layer to extract both local and global channel information. The outputs of both blocks are added, and a sigmoid activation function is applied to them to produce attention weights, which can be defined as:

$$A_c = \lambda\left[F_c\left(A_p\left(\hat{Y}\right)\right) + F_c\left(M_p\left(\hat{Y}\right)\right)\right] \tag{5}$$

Where

$$F_c = W_b * \sigma\left(W_a * Y + b_a\right) + b_b \tag{6}$$

Here, λ represents the sigmoid activation function, A_p and M_p represent average and max pooling, W_a and W_b represent the weight matrices of the convolutional layers, and b_a and b_b are their respective bias terms.

These attention weights are then multiplied with the original feature map to enhance the important features while suppressing the less important ones. Mathematically, it can be defined as:

$$Y_c = A_c \times \hat{Y} \tag{7}$$

These channel-wise refined features are then passed to the spatial attention module to emphasize spatially important locations. Here, the feature map is passed through average and max pooling layers simultaneously to capture both average and max features. The outputs of both layers are then concatenated and passed through a 7×7 convolutional layer and a sigmoid activation function to produce a spatial feature map, which is then used to weigh the spatial features according to their importance. Mathematically, it can be represented as:

$$A_s = \lambda\left(W_s * \left[A_p(Y_c) \parallel M_p(Y_c)\right] + b_s\right) \tag{8}$$

Here, $\parallel$ represents concatenation. This weighted feature map is then multiplied with the channel-wise recalibrated feature map to generate a channel and spatial-wise emphasized feature representation:

$$Y_s = A_s \times Y_c \tag{9}$$

Here, Y_c and Y_s represent the outputs of the channel and spatial attention modules, respectively.

Multi-scale Feature Learning: The recalibrated feature map is then passed to the multi-scale feature extraction block, which consists of four parallel layers, each with a different kernel size to extract information at various scales. Each branch is composed of a convolutional layer, batch normalization, and a ReLU activation function to extract and normalize features.

In Branch 1, the convolutional layer has a kernel size of 1×1 to capture pointwise features while reducing the dimensionality of the feature map. In Branch 2, the kernel size is increased to 3×3 to capture small and mid-sized tumors, while in Branch 3, a convolutional layer with kernel size 5×5 is used to extract contextual information.

The last branch, Branch 4, consists of a max pooling layer to summarize spatial information, followed by a 1×1 convolutional layer, batch normalization, and a ReLU activation to enhance spatial robustness in the model. The outputs of all the branches are concatenated before being passed to the next block. Mathematically, this operation can be represented as:

$$Y_M = \text{concat}[Y_1, Y_2, Y_3, Y_4] \tag{10}$$

Where the output of each branch is calculated as:

$$Y_i = \sigma\left(\beta\left(W_i * Y_s\right) + b_i\right), \quad \text{for } i \leq 3 \tag{11}$$

$$Y_i = Y_i(M_p), \quad \text{for } i = 4 \tag{12}$$

Here, W_i and b_i denote the weights and bias of the i-th convolutional layer, β is the batch normalization function, σ is the ReLU activation, M_p is max pooling, and Y_s is the input from the spatial attention module.

These multi-scale extracted features are then passed through another Mixture of Expert block with same architecture as before, however the depth of network is increased to 256 channels. This block ensures that features are extracted at various level of abstractions and model can learn to handle diverse range and types of tumors. Another channel and spatial attention module is incorporated after this Moe block to remove the noise from feature map and to pay attention on more tumor relevant features. It is then followed by another set of Moe block and attention modules incorporated in the very same architecture but with higher depth to extract progressively complex features. So, the input tensor after passing through three Moe blocks followed by attention modules, finally headed towards the main classifier for final classification.

Final and Auxiliary Classifier:

In the main classifier, the feature map is first passed through an adaptive average pooling layer to summarize the spatial information, and the output is passed to a dense layer that projects this information from a higher-dimensional feature space to a lower-dimensional feature space. A ReLU activation function is applied afterward to introduce non-linearity, followed by a dropout layer with a dropout rate of 0.5, which randomly deactivates 50% of the neurons to prevent the model from overfitting. Finally, another dense layer is used to convert the feature maps into class logits. Mathematically, this can be defined as:

$$C = W_2^{'} \cdot \partial\left(\sigma\left(W_1^{'} \cdot A_p(Y)\right)\right) \tag{13}$$

Here, ∂ represents the dropout layer, and $W_1^{'}$ and $W_2^{'}$ represent the weights of the dense layers.

Apart from the main classifier, two auxiliary classifiers are also introduced in the model—one after the first spatial attention module and another after the second spatial attention module. These auxiliary classifiers predict class labels from intermediate features, helping the model improve feature learning during the early stages of training.

Each auxiliary classifier consists of an adaptive average pooling layer followed by a dense layer, a ReLU activation, a dropout layer, and another dense layer—configured in the same way as the main classifier. The auxiliary losses are added to the model's overall loss only during training. During model evaluation (testing), only the loss from the main classifier is considered.

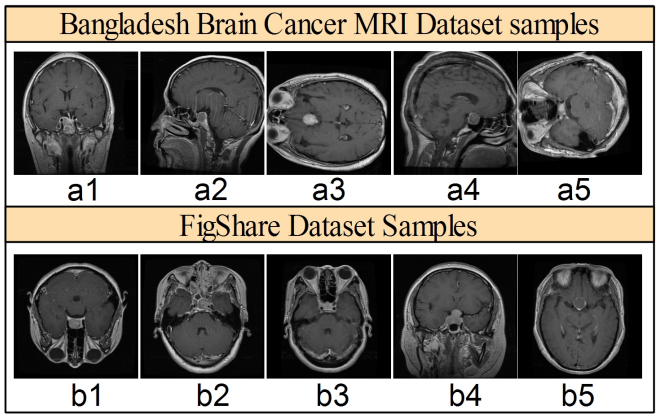

Fig. 4. a1-a5 presented the samples of Bangladesh brain cancer MRI and b1-a5 presented the sample images of Figshare dataset.

IV. Dataset Collection and Augmentation

The experimental design of this work uses two publicly available datasets of brain tumor image classification: https://data.mendeley.com/datasets/mk56jw9rns/1 and https://figshare.com/articles/dataset/brain_tumor_dataset/1512427. In both the Brain Cancer raw MRI data and the Figshare brain tumor dataset, which are displayed in Figure 4, there are three classes in the image classification data for brain tumors. The first dataset contains *Brain_Glioma*, *Brain_menin*, and *Brain Tumor*, while the second dataset includes *Meningioma*, *Glioma*, and *Pituitary Tumor*.

Consequently, Figshare dataset has limited number of images, which are insufficient to train deep learning models effectively. To solve this problem, the original images are splitted into 55% for training, 15% for validation, and 30% for testing. After that,we employed the data augmentation technique on training portion. In this method, three image translations are performed such as flip left, flip right, and rotation to increase the diversity of the dataset. The whole description of the datasets are described in Table I.

TABLE I
Brain tumor datasets with number of classes, images, and splits

No.of Class	Original Images	Train:Test: Validation
Bangladesh Brain Cancer - MRI dataset		
Brain_Glioma	2004	1102 / 602 / 300
Brain_Menin	2004	1102 / 602 / 300
Brain_tumor	2048	11026 / 615 / 307
Figshare brain Tumor dataset		
Glioma	1426	784 / 429 / 213
Meningioma	708	784 / 213 / 106
Pituitary tumor	930	784 / 280 / 139

V. Experimental Setup and Evaluation

A. Training and Parameters

For the training of proposed Mutliscale MoE model, 55% of data is used for training, 15% for validation during the learning phase, and the remaining 30% is allocated for testing. The training process is configured with a mini-batch size of 16, a learning rate of 0.00001, epochs is 50, number of experts is 5, auxiliary weights is 0.03, and the Adam optimizer. To prevent overfitting, early stopping is applied with a learning rate decay factor of 0.2, a patience value of 5 epochs, and a minimum learning rate threshold of 0.00001. All experiments are conducted using Python in a PyTorch environment, executed on a desktop system equipped with an NVIDIA RTX 3060 GPU 12 GB and 24 GB of system RAM.

B. Evaluation Metrics

To rigorously assess the performance of the proposed multiscale MoE model, we based on a comprehensive range of standard and complex evaluation metrics. The standard metrics are Precision, Recall, F1 score, and Support, which together provide a solid assessment of the models discriminative ability, sensitivity, and assessment of classification accuracy. In addition, we incorporated complex metrics such as Balanced Accuracy, F-beta score, Expected Calibration Error (ECE), Brier Score, and an Overall Testing. These metrics provide a fairly complete review of a models efficacy, including dealing with class imbalance, evaluating the tradeoff between precision versus recall when tailoring importance weights, and exploring the calibration of the predicted probabilities. Balanced Accuracy is important in providing equal performance for all classes, no matter how prevalent they are, while the F-beta score provides adjustable weighting so inferences can emphasize recall according to the requirements of any particular application. ECE and Brier Score are important measures of probabilistic reliability, measuring the covariance between predicted certainty and actual outcomes. Calibration-aware metrics are very important in clinical decision making, especially when it comes to high-stakes environments, when both predictive certainty and balanced performance for all classes are important to achieve accurate diagnosis, optimal treatment planning, and ultimately improve overall patient care. the complex metrices are mathematically defined as:

$$\text{BA} = \frac{1}{C}\sum_{i=1}^{C}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{14}$$

$$F_\beta = (1+\beta^2)\cdot\frac{\text{Precision}\cdot\text{Recall}}{(\beta^2\cdot\text{Precision})+\text{Recall}} \tag{15}$$

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{16}$$

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\,|\text{acc}(B_m) - \text{conf}(B_m)| \tag{17}$$

Where C is the number of classes, β calculates the weighting between precision and recall, and N is the number of samples.

VI. Results of Proposed Multiscale MoE

The classification performance of the proposed Multiscale MoE model on the Figshare dataset shows excellent discriminative performance for all three brain tumors, a shown in Table II. Class-wise results indicate that pituitary tumor is classified with the highest precision which is 0.99, demonstrating that while the model correctly classified almost every tumor type, it had an excellent ability to correctly avoid false positives for pituitary tumor while maintaining a high sensitivity to pituitary tumor classification. Glioma classification similarly showed robust performance across both precision and recall with a score of over 0.97 and culminating in an F1-score of 0.98., Meningioma classification while still showing strength precision is0.92, recall is 0.93, and F1-score is 0.93 showed a greater degree of incorrect classification then the other tumors. Examination of the confusion matrix in Figure 5 shows that of the 76 meningioma cases, 12 were found to be glioma and 3 were classified as pituitary tumors. Moreover, of the 59 gliomas, 8 were misclassified as meningiomas. These errors are likely a result of the similarity of radiological appearance for these tumor subtypes, particularly when tumors have overlapping anatomy, as well as more subtle similarities in texture and intensity data that could challenge the MoE model's ability to create distinct class separation boundaries from the MRI scans of these tumor types.

Meanwhile, compared to the Bangladesh dataset in Table II, the proposed model provides greater consistency and balance in performance for each tumor class overall, with each tumor class exhibiting precision, recall and F1-score values exceeding 0.95. Glioma detection remains highly superior precision is 0.98, recall is 0.98, and F1-score is 0.98, meningioma detection achieves balanced precision which is 0.96 and recall is 0.95, pituitary tumor showed exceptional recall is 0.99 with precision of 0.96. In Figure 6, the confusion matrix shows few false predicted classifications with meningioma 10 samples predicted as glioma and 22 samples predicted as pituitary tumor. In addition, a few pituitary tumor such as 9 cases were predicted as meningioma. Generally, the confusion matrix shows fewer cross-class prediction errors when compared to the Figshare results which could be attributed to differing greater inter-class separability. Nevertheless, as can be concluded from both results, the few predicted cases can be easily explained by the overlapping tumor morphology, similar distributions of partial volume effects related to the MRI acquisition.

The extensive performance evaluation of the proposed Multiscale MoE, on the Figshare and Bangladesh datasets, supports the models apparent classification performance, robustness and likely generalization from several sources of data (Figshare and Bangladesh). For the Balance Accuracy measure which is useful to evaluate overall performance across all classes independent of sample distribution was shown to achieve 0.958 for Figshare and 0.9691 in Bangladesh. Both values highlights that the model consistently provides high recognition accuracy independent, or with major class imbalances. Similarly, the F-beta scores that coincidentally achieved the same values (0.9591 for Figshare and 0.9691 for Bangladesh) supports the model's performance aspect of maintaining a fair balance between positive predictive value

TABLE II
Classification results of proposed multiscale MoE on both selected datasets

Class	Precision	Recall	F1-score	Support
Figshare Results				
Glioma	0.97	0.98	0.98	429
Meningioma	0.92	0.93	0.93	213
Pituitary Tumor	0.99	0.96	0.98	280
Bangladesh Results				
Glioma	0.98	0.98	0.98	602
Meningioma	0.96	0.95	0.96	602
Pituitary Tumor	0.96	0.99	0.97	615

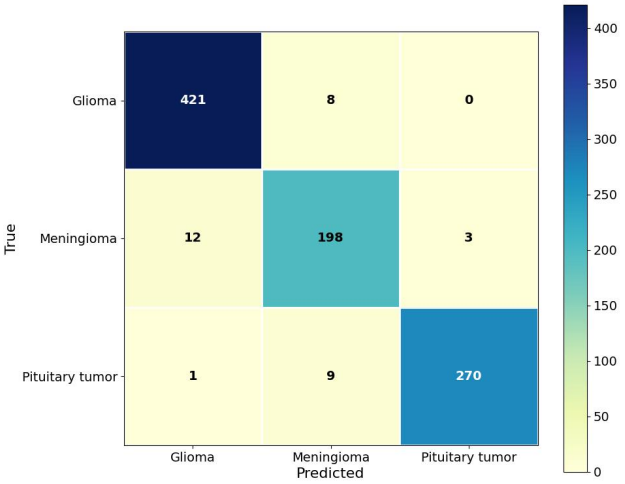

Fig. 5. Confusion matrix of proposed multiscale MoE on Bangladesh brain cancer dataset.

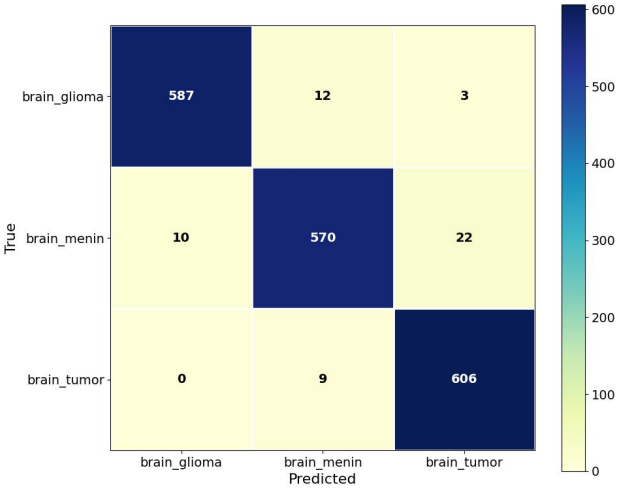

Fig. 6. Confusion matrix of proposed multiscale MoE on Figshare dataset.

(PPV) and sensitivity on both datasets. With the clinically important aim in not creating potential false negatives in brain tumours, we believe the model's F-beta performance being high shows that the model was successfully able to prioritize sensitivity for precision at some gotten trade-off.

Not surprisingly on reliability predictors, both the Expected Calibration Error (ECE) were quite low at 0.0229 for Figshare (and quite low at 0.0086 for Bangladesh) suggestively showing that close the predicted probability provided by model is to the actual observed outcome (the Bangladesh suggesting near perfect calibration). The Brier Score which is a metric of the mean squared difference between predicted probabilities (prediction) and actual outcomes (ground truth), were again shown to be lower in the Bangladesh dataset (0.0456) than Figshare dataset (0.0617), suggesting that model was making more confident and accurate probability assessments. The Overall Testing accuracy achieved over 96% for both datasets (96.42% for Figshare dataset and 96.92% for Bangladesh dataset) and demonstrated good fit and discriminative ability of the model supports its adaptable nature across datasets.

In sum, the performance results all imply that the proposed MoE model is able to provide excellent accuracy in classification, together with probability calibration which is very important in clinical environments where diagnostic accuracy remains paramount while confidence in predicted probabilities play a large role to reliable decision making and effective planning of the treatment decision.

TABLE III
COMPLEX METRICES PERFORMANCE OF PROPOSED MULTISCALE MOE ON SELECTED DATASETS

Metrics	Figshare Results	Bangladesh Results
Balanced Accuracy	0.958	0.9691
F-beta Score	0.9591	0.9691
ECE	0.0229	0.0086
Brier Score	0.0617	0.0456
Overall Testing	96.42	96.92

VII. ABLATION STUDIES

A. Model Computational Complexity

In the first experiment, an ablation study has been conducted aomng the proposed multiscale MoE model against state-of-the-art architectures such as Swin-T, SMViTv2-H and V-MoE based on parameters, GFLOPs and inference time, as shown in Table IV. The comparative results show that although SMViTv2-H [22] has the largest number of parameters 667M and GFLOPs 120.6, it has the highest inference time which is 24.50 seconds. Next, V-MoE [23] has the largest number of parameters which is 14.7B but a moderate GFLOPs value 12.35 and a reasonable inference time which is 11.71 sec, showcasing the efficiency of the expert-based architecture. Given the size of the model, while [24] had the least number of parameters which is 28M and GFLOPs 4.5, its inference time of 12.08 seconds may indicate a smaller representational capacity based on the reduced computational load. Our proposed multiscale MoE model returns a decent trade-off, remaining a lightweight model 25.27 million parameters with competitive computation complexity which is 11.95 GFLOPs. Furthermore, the proposed model exhibits the fastest inference time which is 10.85 sec of the entire study which we believe indicates that the proposed model is able to use multiscale feature processing and expert routing effectively, enabling the delivery of insights quickly and efficiently in a real-time enviroment.

TABLE IV
COMPARISON OF SOTA AND PROPOSED MODEL BASED ON PARAMETERS, GFLOPS, AND INFERENCE TIME.

Models	No. of Parameters	GFLOPS	Inference Time (s)
Swin T [24]	28	4.5	12.08
SMViTv2-H [22]	667M	120.6	24.50
V-MoE [23]	14.7B	12.35	11.71
Proposed	25.27	11.95	10.85

B. Comparison with state of the Art Models

The second ablation study described compares proposed multiscale MoE with leading state-of-the-art models, as shown in Table V. In performance measure F-beta, the proposed multiscale MoE model had the highest score for F-beta which is 0.951 indicating that proposed model provided the most focused balance between precision and recall score, and was best suited for classification tasks where the cost of both false positives and false negatives are critical. The ECE score of proposed model 0.0223 was slightly worse than the SMViTv2-H [22] model, which is 0.018, indicating that while our model was simultaneously reliable, SMViTv2-H [22] had a very slightly better measure of confidence with the prediction probability. Importantly, For the Brier Score the MoE had totally lowest value calculated for the Brier Score which is 0.0603 indicating that our proposed model had the highest accuracy of probability prediction. Finally, in testing accuracy score, our proposed MoE model had the highest score of 96.14% compared to all SOTA methods. While each individual performance measure demonstrates that the proposed MoE outperformed some of the competing methods, the collective measures prove that our proposed model not only had predictive performance better than state-of-the-art models, but also more reliable prediction probability and estimates.

TABLE V
ABLATION STUDY BASED ON F-BETA, ECE, BRIER SCORE, AND ACCURACY AMONG THE PROPOSED AND SOTA METHODS

Methods	F-beta	ECE	Brier Score	Accuracy (%)
Swin T [24]	0.925	0.030	0.075	94.80
SMViTv2-H [22]	0.945	0.018	0.062	95.90
V-MoE [23]	0.935	0.025	0.067	95.40
MViTv2-B [25]	0.940	0.020	0.065	95.60
Proposed MoE	0.951	0.0223	0.0603	96.14

C. Expert Utilization Across Selected Datasets

In this experiment, the utilization of each experts of proposed mutilscale MoE has been measured across the bangladesh brain cancer dataset, as shown in Figure 7. this

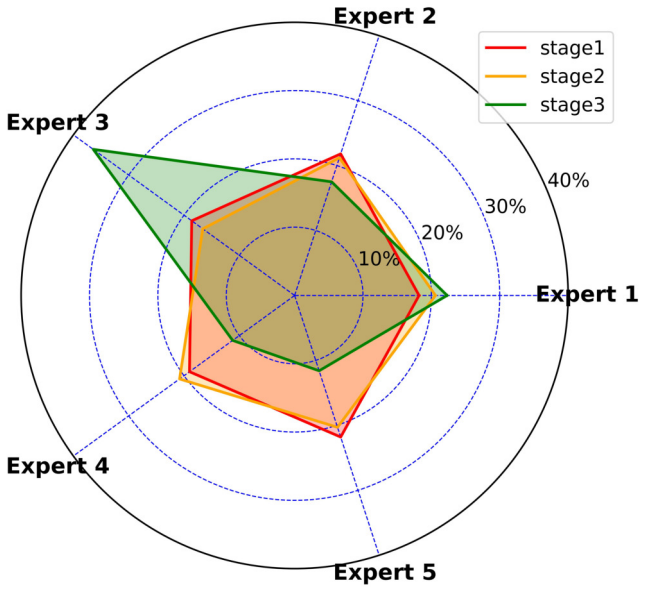

Fig. 7. Expert utilization patterns across the three stages of the proposed Multiscale Mixture of Experts (MoE) model on the Bangladesh Brain Cancer MRI dataset. The radar plot shows the percentage of routing weight assigned to each expert (Expert 1–Expert 5) by the dynamic router during different processing stages. In stages 1 and 2 (red and orange lines), the routing distribution is relatively balanced across experts, with moderate emphasis on Experts 1, 2, 4, and 5. In stage 3 (green line), Expert 3 dominates with nearly 40% routing weight, indicating its specialization in extracting high-level, fine-grained features critical for final classification, while other experts contribute in supporting roles.

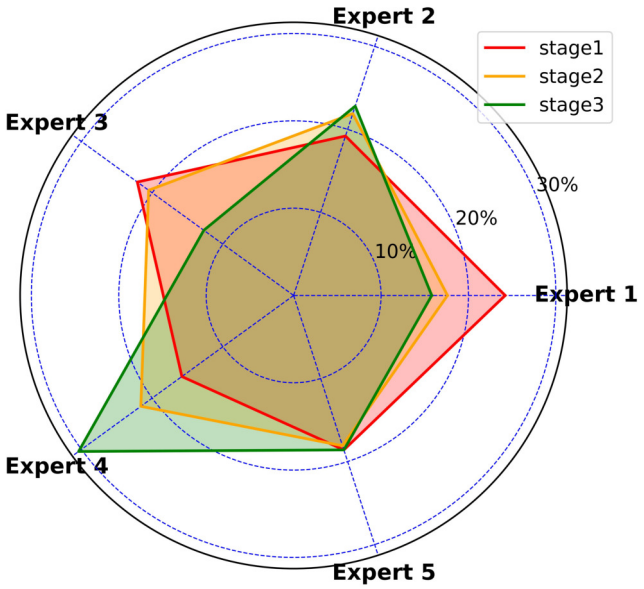

Fig. 8. Expert utilization patterns across the three stages of the proposed Multiscale Mixture of Experts (MoE) model on the Figshare Brain Tumor dataset. The radar plot displays the percentage of routing weight assigned to each expert (Expert 1–Expert 5) by the dynamic router during stage 1 (red), stage 2 (orange), and stage 3 (green). While stages 1 and 2 show a relatively balanced distribution across experts with moderate emphasis on Experts 2, 3, and 4, stage 3 exhibits a strong shift toward Expert 4, which receives the highest routing weight (over 30%), indicating its specialization in extracting discriminative features for the Figshare dataset's unique imaging characteristics.

plot clearly illustrates a distinct usage pattern of the proposed multiscale MoE experts in three stages. During the early stages, Experts 1, 2, 4, and 5 have relatively balanced and moderate routing weights, demonstrating the use of a distributed feature extraction approach in the beginning stages. A major shift occurred in stage 3 when the routing weight of Expert 3 increased substantially, to nearly 40%. This routing weight reflects that Expert 3 was the most prominent expert contributing to the eventual prediction decision. The increased weight of Expert 3 suggests that it was especially proficient at producing the fine-grained, high-level features required to classify thesubtle cases within the Bangladesh dataset. The other experts clearly contributed during this stage but their routing weights had decreased considerably, suggesting they were only being used in a supportive role of general feature representation rather than specific discrimination. The reliance on Expert 3 at the final stage represents the most low-level decision making and should provide a perspective about the versatility of the proposed model where the routing mechanism initiated relevant extremely relevant experts depending on the specific feature characteristics of the dataset.

In Figshare dataset, the routing behavior of experts shows an opposing pattern in specialization. The most apparent observation in this context was the increase in Expert 4 usage in stage three, overtaking Expert 3 whose share dropped significantly, as shown in Figure 8. Expert 4 became the expert with the highest routing weight during stage 3, with over 30%. This suggests that Expert 4 has learned feature extraction that is more appropriate given the unique imaging features of the Figshare dataset, which could include variations in tumor boundaries, texture distributions, and contrasts specific to different imaging modalities. All other experts 1, 2 and 5 showed relatively stable routing weights across the staging, indicating that the same experts participated consistently and contributed relatively equally in the decision making process when comparing with the Bangladesh dataset. Expert 3 lost its influence by later stages; it is clear from the data that what it had learned were less informative features for high level refinement processes in Figshare brain tumor model. Overall, the Figshare results exhibit more evenly shared expert utilization, where it appears the model was able to shift the use of experts on-the-fly as required; dynamically shifting emphasis under the final stage to Expert 4 when classifying images based on this dataset's unique properties.

VIII. Conclusion

Brain tumors are a significant global health challenge because of their inherent aggressiveness, poor patient prognosis, and difficulty in diagnosis. Notably, MRI scanning is acknowledged as the gold-standard diagnostic modality and ongoing treatment tool for brain tumors; however, automated classification techniques often struggle with the highly variable

morphologies of brain tumors, overlap in features, and different conditions on images. In this work, the study proposes multiscale MoE, a Mixture of Experts framework that unifies a CNN and attention-based modules in a single model that support multiscale approaches. The framework builds into the architecture a way to efficiently capture tumor-specific features using multiscale feature extraction, channel-spatial attention, and dynamic expert routing mechanism.

This study used two public datasets: the Bangladesh Brain Cancer MRI database and Figshare Brain Tumor dataset. the presented framework achieved precision, recall, and F1-scores greater than 0.95 for all classes of tumors including a balanced accuracy of 0.9691, a F-beta score of 0.9691, an ECE of 0.0086, a Brier score of 0.0456, and a general overall accuracy of 96.92%. With the Figshare dataset, precision scores approached 0.99, recalls approached 0.98, the balanced accuracy score was 0.958, the F-beta score approached 0.9591, the ECE was 0.0229, the Brier score was 0.0617, and a general overall accuracy of 96.42%. Compared to other state of the art models and approaches, multiscale MoE produced the highest testing accuracy 96.14% and the lowest Brier score 0.0603, and was ultimately the fastest model demonstrating inference times of 10.85 s without tuning.

Despite the promising results, there are limitations with the proposed multiscale model. it is reliant on MRI data because it was only trained and validated with data from two publicly accessible datasets and may not sufficiently reflect the vast variability routinely encountered in imaging protocols, tumor subtypes, and demographics when treating patients globally. This may have implications on the model's generalizability when implemented and applied in various medical environments.

Future work will focus on more data from more than one institute and add additional imaging modalities such as with PET and CT so that the model may learn from more various characteristics in tumours. Domain adaptation and transfer learning methods can be used to improve robustness across scanners and imaging acquisition protocols.

REFERENCES

[1] S. Lapointe, A. Perry, and N. A. Butowski, "Primary brain tumours in adults," *The Lancet*, vol. 392, no. 10145, pp. 432–446, 2018.

[2] D. Sipos, B. L. Raposa, O. Freihat, M. Simon, N. Mekis, P. Cornacchione, and Á. Kovács, "Glioblastoma: clinical presentation, multidisciplinary management, and long-term outcomes," *Cancers*, vol. 17, no. 1, p. 146, 2025.

[3] J. Huang, H. Li, H. Yan, F.-X. Li, M. Tang, and D.-L. Lu, "The comparative burden of brain and central nervous system cancers from 1990 to 2019 between china and the united states and predicting the future burden," *Frontiers in Public Health*, vol. 10, Oct. 2022.

[4] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," *Ca*, vol. 75, no. 1, p. 10, 2025.

[5] A. Moiyadi, V. Singh, R. Tonse, and R. Jalali, "Central nervous system (cns) tumors," in *Tata Memorial Centre Textbook of Oncology*, pp. 379–404, Springer, 2024.

[6] M. Ijaz, I. Hasan, B. Aslam, Y. Yan, W. Zeng, J. Gu, J. Jin, Y. Zhang, S. Wang, L. Xing, *et al.*, "Diagnostics of brain tumor in the early stage: current status and future perspectives," *Biomaterials Science*, 2025.

[7] S. Gunasekaran, P. S. Mercy Bai, S. K. Mathivanan, H. Rajadurai, B. D. Shivahare, and M. A. Shah, "Automated brain tumor diagnostics: Empowering neuro-oncology with deep learning-based mri image analysis," *Plos one*, vol. 19, no. 8, p. e0306493, 2024.

[8] A. Nazir, A. Hussain, M. Singh, and A. Assad, "Deep learning in medicine: advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis," *Multimedia Tools and Applications*, vol. 84, no. 17, pp. 17677–17740, 2025.

[9] R. Disci, F. Gurcan, and A. Soylu, "Advanced brain tumor classification in mr images using transfer learning and pre-trained deep cnn models," *Cancers*, vol. 17, no. 1, p. 121, 2025.

[10] V. Hassija, B. Palanisamy, A. Chatterjee, A. Mandal, D. Chakraborty, A. Pandey, G. Chalapathi, and D. Kumar, "Transformers for vision: A survey on innovative methods for computer vision," *IEEE Access*, 2025.

[11] D. Hussain, M. A. Al-Masni, M. Aslam, A. Sadeghi-Niaraki, J. Hussain, Y. H. Gu, and R. A. Naqvi, "Revolutionizing tumor detection and classification in multimodality imaging based on deep learning approaches: Methods, applications and limitations," *Journal of X-Ray Science and Technology*, vol. 32, no. 4, pp. 857–911, 2024.

[12] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," *Technologies*, vol. 13, no. 1, p. 32, 2025.

[13] S. Tehsin, I. M. Nasir, R. Damaševičius, and R. Maskeliūnas, "Dasam: Disease and spatial attention module-based explainable model for brain tumor detection," *Big Data and Cognitive Computing*, vol. 8, p. 97, Aug. 2024.

[14] S. Tehsin, I. M. Nasir, and R. Damaševičius, "Gatransformer: A graph attention network-based transformer model to generate explainable attentions for brain tumor detection," *Algorithms*, vol. 18, p. 89, Feb. 2025.

[15] C. Srinivas, N. P. KS, M. Zakariah, Y. A. Alothaibi, K. Shaukat, B. Partibane, and H. Awal, "Deep transfer learning approaches in performance analysis of brain tumor classification using mri images," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 3264367, 2022.

[16] H. A. Khan, W. Jue, M. Mushtaq, and M. U. Mushtaq, "Brain tumor classification in mri image using convolutional neural network," *Mathematical Biosciences and Engineering*, 2021.

[17] A. Pashaei, H. Sajedi, and N. Jazayeri, "Brain tumor classification via convolutional neural network and extreme learning machines," in *2018 8th International conference on computer and knowledge engineering (ICCKE)*, pp. 314–319, IEEE, 2018.

[18] K. Muhammad, S. Khan, J. Del Ser, and V. H. C. De Albuquerque, "Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 507–522, 2020.

[19] C. Narmatha, S. M. Eljack, A. A. R. M. Tuka, S. Manimurugan, and M. Mustafa, "A hybrid fuzzy brain-storm optimization algorithm for the classification of brain tumor mri images," *Journal of ambient intelligence and humanized computing*, pp. 1–9, 2020.

[20] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, and S. W. Baik, "Multi-grade brain tumor classification using deep cnn with extensive data augmentation," *Journal of computational science*, vol. 30, pp. 174–182, 2019.

[21] H. Mzoughi, I. Njeh, A. Wali, M. B. Slima, A. BenHamida, C. Mhiri, and K. B. Mahfoudhe, "Deep multi-scale 3d convolutional neural network (cnn) for mri gliomas brain tumor classification," *Journal of Digital Imaging*, vol. 33, no. 4, pp. 903–915, 2020.

[22] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814, 2022.

[23] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8583–8595, 2021.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[25] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814, 2022.