Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Generative AI for Early Disease Detection: Hybrid ViT-CNN with Multi-Head Attention in Medical Imaging

Raiyan Muntasir Monim[1], Kamrul Islam[2,*], Sabit Md Asad[1], MD Ahbab Hussain[3], Belayet Hossen[4], Md Takbir Alam Manjar[4], Sharmin Sultana[5]

[1]College of Graduate and Professional Studies, Trine University, Angola, IN 46703, USA

[2]Gabelli School of Business, Fordham University, Bronx, NY 10458, USA

[3]Ketner School of Business, Trine University, Angola, IN 46703, USA

[4]Master of Science in Information Systems Management, Stanton University, Anaheim, CA 92802, USA

[5]College of Science & Mathematics, Rowan University, Glassboro, NJ 08028, USA

[*]Corresponding Author E-mail: ki7@fordham.edu

ARTICLE INFO

Received: 08 Jan 2026

Revised: 22 Feb 2026

Accepted: 28 Feb 2026

ABSTRACT

Early recognition of disease in medical imaging gives a good chance for fast treatment and an increase in survival in serious cases like brain tumors. To better classify brain tumors from MRIs, our study introduces a new type of deep learning method that combines CNNs and ViT for improved results. Though strong at finding out small features in an image, traditional CNNs fail to notice big relationships in the image because their receptive fields are narrow. Instead, ViT brings in multi-head self-attention to allow the model to focus on long interactions between different parts of the image. By mixing the two types of architectures, this research uses CNNs to zoom in and ViTs to look at the full picture. The study makes use of a brain tumor MRI dataset that anyone can access for free from Kaggle. Normalization, resizing, and augmentation methods were all utilized to increase the model's strength and ability to generalize. A stratified 80-20 data split was used to develop and verify the hybrid model. The clear boost suggests that adding local detail perception from CNN and the global influence of ViT can help our model better interpret medical images. Adding Grad-CAM showed us which parts of the scans were the most important to the decision made by the model, making the overall behavior of the model more obvious. Overall, this work establishes that CNN-Transformer architectures are helpful for medical imaging and prepares the way for introducing AI-assisted diagnostics in clinical practice to aid faster choices.

Keywords: Brain Tumor, CNN, ViT, Medical Imaging, Computer-aided Diagnosis

1. INTRODUCTION

Deep learning (DL) and computer vision technologies have gained prominence in medical diagnostics, particularly in the identification of brain tumors via MRI scans [1]. Advancements in medical technology facilitate timely and accurate diagnoses. This yields an efficacious therapeutic strategy that enhances patient outcomes [2]. Radiologists and clinical histopathologists are increasingly dependent on medical imaging data to enhance diagnostic precision [3]. The substantial quantities of medical records and the widespread adoption of digital imaging technologies have established a critical function for artificial intelligence (AI)-driven automated decision support systems. Utilizing these technologies to analyze substantial data volumes facilitates prompt and precise evaluation of medical pictures, thus enhancing clinical effectiveness [4].

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The Convolutional Neural Networks (CNN) model has exhibited proficiency in detecting brain tumors in MRI scans and identifying lung pathology in chest X-rays [5]. Although CNNs effectively capture local information, their limited receptive fields constrain their capacity to encompass global context. This shortcoming impedes their performance on increasingly complicated tasks, such as categorizing brain tumors, which necessitate comprehension of the relationships between distant portions of an image [6]. To address the limitations of CNNs, Vision Transformers (ViTs), which utilize transformer-based architectures to enhance multi-head self-attention for capturing long-range dependencies in images, have been recently developed [7]. It is observed how ViTs can acquire global context from medical images [8]. The capacity to comprehend long-range relationships is essential when addressing complex patterns [9]. Despite their success, ViTs neglect local spatial cues, which are crucial for accurate diagnosis . Integrating the advantages of CNNs and ViTs to attain both local and global perception presents a difficulty in the development of stable hybrid models proficient in recognition tasks [10]. While other fields have explored hybrid models combining CNN and ViT, the use of such approaches to brain tumor diagnosis by MRI remains underexplored. The existing gap in research drives the current investigation. The research seeks to create an innovative hybrid ViT-CNN model capable of enhancing brain tumor categorization. The primary objective of this study is to develop a hybrid CNN-ViT model for the categorization of brain tumors in MRI images. The research will investigate the integration of CNNs, proficient in local feature extraction, with ViTs, which excel in capturing long-range global dependencies in images. We will create a hybrid model by integrating CNNs with ViTs, resulting in enhanced classification accuracy and improved interpretability compared to either CNNs or ViTs independently.

The objective of this research is to enhance the classification model's reliability and generalizability by data augmentation, scaling, and normalization. The strategies are employed to address the variability in the MRI dataset and to enhance the model's operational efficiency across diverse domains. This study is pertinent to AI-assisted brain tumor diagnosis in medical imaging. Accurate and fast detection of a brain tumor is essential for prompt action. Despite advancements in imaging techniques, numerous difficulties persist. This is particularly applicable to intricate cases involving multiple tumor types or when nuanced distinctions are present. The hybrid CNN-ViT model addresses the identified issue by integrating two sophisticated AI architectures to leverage their combined strengths. The researchers who developed the model aspire for their outcomes to be more efficacious in the future. They assert that their methodology will enhance physicians' ability to detect malignant tumors via MRI scans. Moreover, methodologies such as data augmentation and normalization enhanced the model’s generalizability, enabling its application to diverse MRI image datasets in prospective clinical settings.

This architecture has the potential to transform the methodology of brain tumor diagnosis in clinical practice, offering a more dependable and interpretable instrument that facilitates expedited diagnosis and improved health outcomes. The research intends to develop a hybrid ViT-CNN model for the categorization of MRI brain tumors. The method that integrates CNNs with ViTs shows potential for enhancing the precision and interpretability of AI-driven diagnostic models. The following contributions are included in this study:

- Investigating the enhancement of brain tumor classification performance from MRI images by the hybrid integration of CNN and ViT architectures.
- The efficiency of the hybrid model’s generalization by data augmentation, scaling, and normalizing strategies is examined.
- Grad-CAM-based interpretability to enhance the clinical dependability and acceptance of the hybrid model is assessed.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The rest of this paper follows the structure in the following manner. Section 2 presents a literature review of previous research on CNN-, ViT-, and hybrid CNN-Transformer methods of image analysis of brain-tumors with MRI, as well as the main constraints and gaps in the studies. Section 3 describes the dataset, preprocessing and augmentation pipeline, train-test splitting strategy, and the suggested hybrid ViT-CNN architecture. Section 4 presents comparative performance between CNN, ViT, and the proposed hybrid between the proposed hybrid and the three metrics, as accuracy, precision, recall, F1-score, confusion matrices, ROC-AUC, and explainability analysis with Grad-CAM. The findings are interpreted in section 5, which underscores practical implications and limits, and gives guidelines on ways to improve and implement the results into clinical practice. Lastly, Section 6 concludes this study with a summary of the key contributions and findings of the study.

2. LITERATURE REVIEW

Numerous studies have focused on how DL, ViTs, Explainable artificial intelligence (XAI), and hybrid models can aid in diagnosing brain tumor segmentation using MRI, leading to improved diagnostic accuracy and clinical decision-making. The categories into which the available literature is classified are model architectures, application areas, methodological approaches, and constraints. This is followed by the identification of those research gaps so that the proposed hybrid ViT-CNN model can address them. CNNs are widely used in medical imaging analysis for their ability to extract spatial information from magnetic resonance imaging (MRI) scans. For multiclass brain cancer classification, Rasheed et al. [11] designed a CNN model with good accuracy in all tumor categories. Raza et al. [12] observed that global as well as local information can be captured within CNN, and the performance could improve further, particularly when using DenseNet121 in combination with Inception-V2 architectures. Rasool et al. [13] introduced SVMs in hybrid CNN models to enhance the generality and stability of classifications. To enhance the classification accuracy and statistical feature representation, CNNs with several convolutional backbones are desirable. In order to incorporate more features, Nassar et al. [14] proposed a CNN ensemble that is composed of a combination of diverse convolutional architectures. Bansal et al. [15] showed that transferable CNNs can generalize well across heterogeneous datasets, and Biswas and Islam [16] improved generalization by fusing CNNs and SVMs. CNNs are great for some tasks, but ViTs have taken off as efficient models for capturing global dependencies in picture patches. In practice, self-supervised ViTs can outperform CNNs in capturing the global dependencies among image patches. Karagoz, Nalbantoglu, and Fox [17] proposed a self-supervised model named Residual Vision Transformer (ResViT). It can also automatically discriminate tumors with minimal annotation effort. Deeper into promoting the transfer learning, Al-Hamza [18] demonstrated that ViT models can achieve competitive classification accuracy with less training data. An improved ViT model was introduced by Khaniki et al. [19], which improved classification accuracy by feature recalibration and selective cross-attention. Chandraprabha, Ganesan, and Baskaran [20] proposed a hybrid model that fuses the feature representation of CNN and transformer for context modelling. Furthermore, Krishnan et al. [21] proposed a rotation-invariant ViT model to solve the problem of orientation sensitivity in medical imaging, like MRI images.

Using DL techniques for classifying brain tumours could improve diagnostic accuracy in the future. Gasmi et al. [22] proposed CNNs for cancer diagnosis, and the effects of weight initialisation and fusing multiple models on enhancing prediction performance were examined to merge architectures. Aly, Ghallab, and Fathi [23] proposed a ViT-GRU-XAI model for tumour classification. The model combines the ability of CNNs for spatial feature extraction and GRUs for the temporal feature learning. Also, Butt et al. [24] and Rasheed et al. [25] introduced hybrid CNN models, taking as input activation maps of several CNN backbones to improve tumour classification. Crucial for assessing the

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

size and particular stage of a cancer, even larger stages in tumor segmentation have seen remarkable improvements with more sophisticated CNN and ViT models. A 3D U-net-based segmentation model, which integrates spatial and contextual characteristics to demarcate tumours, was presented by Butt and Jabbar [26]. In their description of a segmentation model, Ghazouani, Vera, and Ruan [27] introduced Swin Transformer. Local self-attention is exploited in this model to improve multimodal segmentation performance.

The complementary use of structural, functional, and contrast-enhanced MRI images has increased the precision in tumor segmentation and, as a consequence, multimodal methods have emerged. With missing MRI modalities handling, Kang et al. [28] achieved good results, employing the multimodal feature fusion. A novel 3D multi-cue model for multimodality tumour segmentation was presented by Huang, Chen, and Zhou [29]. Resorting to the self-attention and cross-attention atoms at multiple scales, this model further pushes forward the field of natural language processing. As transfer learning is effective when training data are limited, it has been a popular technique in brain tumor classification. Al-Hamza [18] demonstrated that ViT models trained with transfer learning can achieve good classification performance even using very few labelled data. Mathis-Ullrich and Zeineldin [30] proposed a unified CNN architecture using transfer learning for gliomas' segmentation and paediatric tumors. Today, these models are successful with few annotations since transfer learning is employed for training.

Another fertile territory for a large variety of learning forms is the classification of brain tumours. Ahmed et al. [31] integrated the XAI approach with the ViT-GRU hybrid model to establish an ensemble architecture for brain tumor detection. They improved diagnostic accuracy as well as made a more interpretable forecast. Such a design could employ various pooling techniques in concert with other learning algorithms. A large step forward for translational science, XAI allows physicians to work alongside DL models and better comprehend their findings. Research conducted recently indicates that ViTs have significant potential for the categorization of brain cancers. Through the use of ViTs in ensemble approaches, Tummala et al. [32] demonstrated an improvement in diagnostic accuracy. In order to improve MRI-based tumor classification even further, Khaniki et al. [33] expanded on this notion and used feature calibration and selective cross-attention algorithms. When SHAP is combined with ViT, then the clinical decision-making could be better interpreted, as shown by Tanone, Li, and Saifullah [34]. Table 1 provides a comparative overview of current CNN, Vision Transformer, and hybrid CNN-Transformer-based methods in the brain tumor MRI analysis, their main contributions to the field, as well as their drawbacks.

Table 1 Comparative analysis of existing different DL methods for brain tumor classification

Study	Model	Key Contributions	Limitations/Gaps
Rasheed et al. [11]	CNN	Effective multiclass brain tumor classification with good accuracy across tumor categories.	Limited capability in capturing long-range global dependencies.
Raza et al. [12]	DenseNet121 + Inception-V2	Captures both local and global features; improved classification performance.	Increased model complexity and computational cost.
Rasool et al. [13]	CNN + SVM	Improved generality and stability through hybrid CNNâ€“SVM learning.	Requires careful feature selection; limited scalability.
Nassar et al.	CNN Ensemble	Combines diverse convolutional	High computational

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[14]		backbones for richer feature representation.	overhead; complex training process.
Bansal et al. [15]	Transferable CNN	Strong generalization across heterogeneous datasets.	Performance sensitive to domain shift severity.
Biswas & Islam [16]	CNN + SVM	Enhanced generalization by fusing deep and classical learning models.	Additional tuning required for hybrid optimization.
Karagoz, Nalbantoglu, and Fox [17]	ResViT (Self-Supervised ViT)	Captures global dependencies with minimal annotation effort.	Training complexity; interpretability challenges.
Al-Hamza [18]	ViT + Transfer Learning	Competitive accuracy with limited labeled data.	Still data-hungry for very small datasets.
Khaniki et al. [19]	Enhanced ViT	Improved accuracy via feature recalibration and selective cross-attention.	Increased architectural complexity.
Chandraprabha , Ganesan, and Baskaran [20]	Hybrid CNN-Transformer	Combines CNN spatial features with transformer-based context modeling.	Requires careful feature fusion strategy.
Krishnan et al. [21]	Rotation-Invariant ViT	Addresses orientation sensitivity in MRI images.	Limited validation on diverse datasets.
Gasmi et al. [22]	CNN Ensemble	Improved diagnosis through model fusion and weight initialization.	Computationally expensive ensemble training.
Aly, Ghallab, and Fathi [23]	ViT-GRU-XAI	Integrates spatial, temporal learning with explainability.	Higher training and inference complexity.
Butt et al. [24]	Hybrid CNN	Uses multi-backbone activation maps to improve classification.	Large memory footprint.
Rasheed et al. [25]	Hybrid CNN	Enhanced tumor classification via multi-backbone feature fusion.	Limited explainability.
Butt and Jabbar [26]	3D U-Net	Effective tumor segmentation using spatial and contextual features.	High computational and memory requirements.
Ghazouani, Vera, and Ruan [27]	Swin Transformer	Local self-attention improves multimodal segmentation.	Sensitive to modality quality variations.
Kang et al. [28]	Multimodal Feature Fusion	Handles missing MRI modalities effectively.	Performance degrades with severe modality absence.
Huang, Chen,	3D Multi-Cue Model	Multi-scale self- and cross-	Complex architecture

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and Zhou [29]		attention enhances segmentation.	and training difficulty.
Mathis-Ullrich and Zeineldin [30]	Transfer Learning CNN	Unified framework for glioma and pediatric tumor segmentation.	Limited adaptability to unseen tumor types.
Ahmed et al. [31]	ViT-GRU-XAI Ensemble	Improved diagnostic accuracy and interpretability.	Computationally intensive ensemble learning.
Tummala et al. [32]	Ensemble ViT	Improved MRI-based tumor classification.	Requires large-scale training data.
Khaniki et al. [33]	ViT with Cross-Attention	Enhanced MRI classification via selective attention.	Increased model complexity.
Tanone, Li, and Saifullah [34]	ViT + SHAP	Improved clinical interpretability of tumor classification.	SHAP computation adds overhead.

There have been advances, but multiple limitations in current research continue to make the application of such models difficult in clinical practice. The absence of massive labelled data is a major obstacle in training ViTs and other DL models. These issues have been alleviated to some extent with transfer learning, but their use and applicability are still limited by the use of pre-trained models. One of the big issues is its interpretability. Although there has been progress in XAI-FHVX and the combined use of models, it is still difficult to ensure complete interpretability while keeping high accuracy. If the explanations of AI's decisions are insufficient or if they are irrelevant, clinicians might be skeptical of computerized diagnoses. Many design processors are required, which is one barrier that prevents its application to real-time medical environments. These hybrid models of CNNs and ViTs are computationally expensive and generally impractical to implement in the clinic setting, where rapid real-time output is required.

Finally, as MRI images are orientation-sensitive in nature, architectures like ViTs may not perfectly generalize across brain image orientations. This is critically important because medical images are often acquired in multiple views and orientations. In response to many deficiencies of the previous approaches, we proposed a hybrid ViT-CNN model. Such a hybrid model has strong model generalization ability, efficient processing of complex MRI data, as well as high performance on both tumor segmentation and classification, since it can combine the local feature extraction power of CNNs and the global context analysis ability of ViTs. Evaluating clinicians can have more trust in AI-assisted inferences due to the fact that XAI modules are injected into this model, which certainly facilitates report processing and interpretation. Finally, volumetric imaging techniques such as ViTs and DCNNs have made significant advances for segmenting and classifying brain tumors based on MRIs. However, the use case of CNNs and ViTs in combination has remained unexplored. Integrating these models with XAI may help us address existing limitations and improve diagnostic accuracy, robustness, and clinical utility.

3. METHODOLOGY

This section provides the overall methodology followed in this study regarding the early detection of diseases through medical imaging. The suggested structure is based on the combination of CNN and ViT with an aim of embracing both local feature extraction and global contextual modeling. Figure 1

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

shows the entire process of the research by illustrating the steps implemented at each phase of the data acquisition process to eventual classification and performance analysis. The method entails medical imaging data collection and preprocessing, which comprises normalization and resizing to make them compatible with the deep learning models. The processed images are further input into a three-architecture framework, consisting of a CNN-based architecture, a ViT-based architecture, and a hybrid CNN-ViT architecture, which is the suggested framework. The CNN branch deals with learning fine-grained local spatial features, whereas the ViT branch learns long-range dependencies with the help of multi-head self-attention. In the hybrid structure, the features obtained in each branch are integrated to come up with a complete representation and then classified.

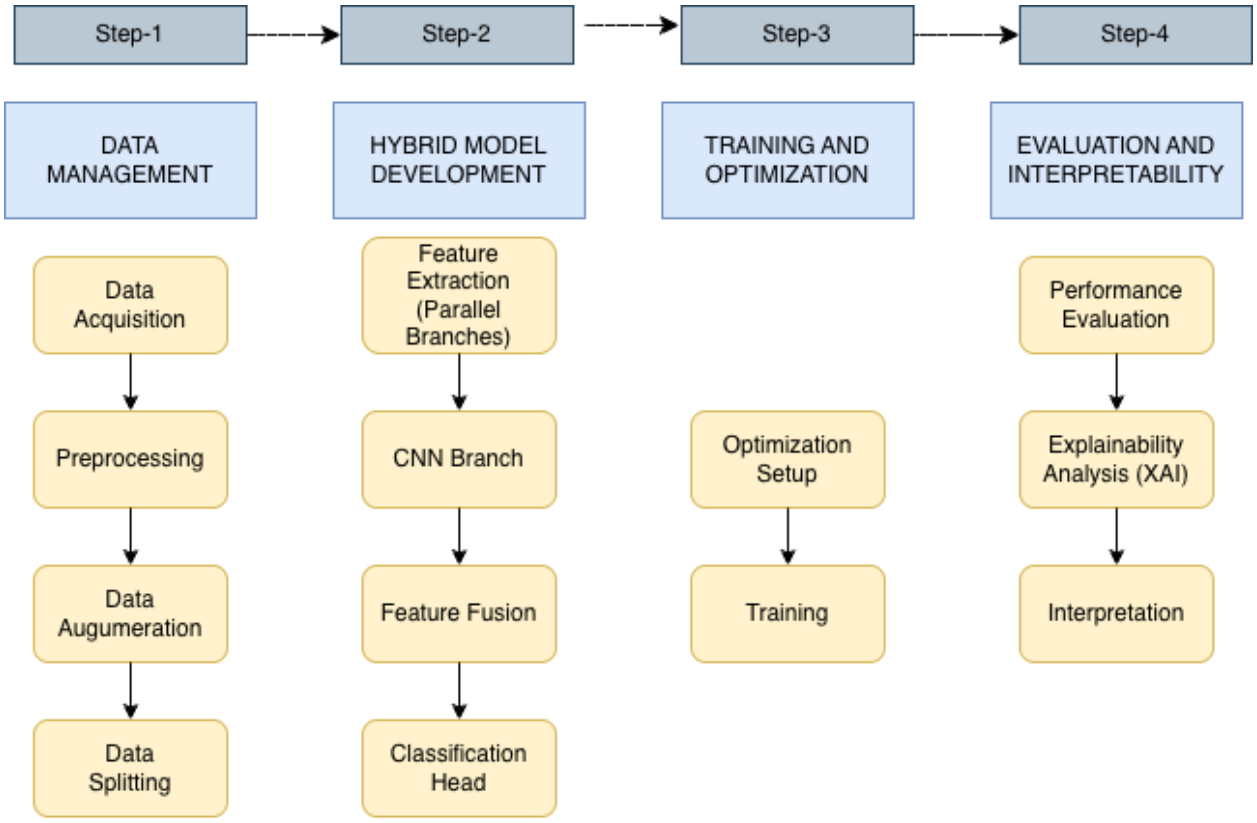

Fig. 1 Workflow for medical image classification using a hybrid CNN-ViT architecture

3.1 Dataset Collection & Description

For this research, an open-access brain tumor dataset of MRI images is used from Kaggle [35]. The data used includes previously obtained and cautiously anonymized MRI scans to respect a patient's confidentiality. None of the information personalizes the data, and the dataset is available only to people using it for study and research. Typically, MRI is used because it is gentle and considered the best procedure for finding changes in the brain. In the dataset, there are 7023 images from T1-weighted contrast-enhanced MRI scans that have been sorted into four classes: glioma, meningioma, pituitary, and no tumor. All images come with a label and are sorted into their proper class for supervised learning to work. The representative images of each type of MRI are shown in Figure 2.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

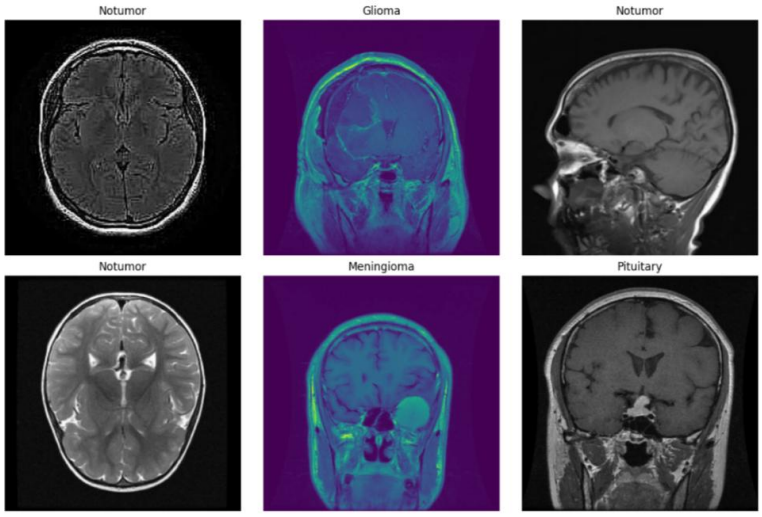

Fig. 2 Sample augmented Images of different categories (Glioma, Meningioma, Pituitary, and No Tumor) of the Brain MRI Image Dataset

3.2 Data Preprocessing

- **Resizing:** All the images that have been explored were RGB images with a common 224×224 pixels resolution. There was a broad range of grayscale values in the tumor pixel histograms, suggesting that the tumors can be identified through contrast, a benefit for DL algorithms. This ensures that it simultaneously preserves important structural features and reduces computational complexity.
- **Rescaling:** Pixel values were adjusted by applying a rescale factor of 1/255, normalizing them to the interval [0, 1]. This improves numerical stability and expedites model convergence by alleviating substantial gradient variations.
- **Image data augmentation**: In medical imaging (especially in MRI images), augmentation of data is essential. Magnetic field distortions can be many and come about due to many factors as a result of patient motion, magnetic field irregularities, and equipment failures. The data augmentation can address such issues by generating more synthetic images using the original MRI images. In the case of MRI imaging, the customary data augmentation techniques are intensity normalization, flipping, rotation, and scaling. Figure 1 presents some augmented images.

3.3 Data Splitting and Validation

A thorough EDA study was carried out to learn how the dataset is organized and what its distribution looks like. A total of 7023 T1-weighted contrast-enhanced MRI images were used in the study, organized into four classes: glioma tumor (1321 images), meningioma tumor (1339 images), pituitary tumor (1457 images), and notumor (1595 images). This data distribution is moderately unbalanced due to a small number of notumor samples. It is necessary to solve this issue to prevent bias when evaluating the outcome of classification. Several methods to process and increase the training data were applied before training and evaluating the model. Before proceeding, the dataset was normalized and resized to create similar pixel values and similar-sized images for all samples. To increase the number of different examples and fight against overfitting, given the imbalance among classes, the data was flipped horizontally and vertically, rotated, zoomed, and shifted.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2 Train-Test Splitting Details of The Brain Tumor Dataset

Dataset Name	Total Samples	Train: Test Ratio	Train Samples	Test Samples
Brain Tumor Dataset	7023 images	80:20	5618	1404

In medical imaging, particularly in MRI images, data augmentation is essential. Magnetic field distortions can be numerous and arise from various factors, including patient motion, magnetic field irregularities, and equipment failures. Data augmentation can address such issues by generating additional synthetic images using the original MRI images. In the case of MRI imaging, the customary data augmentation techniques are intensity normalization, flipping, rotation, and scaling. Figure 2 presents some augmented images.

3.4 Classification Model Architecture & Implementation

For this research, TensorFlow 2.x and Keras have been used to develop and train a ViT and a combined CNN-ViT model for classifying brain tumors. The dataset was loaded from Kaggle ("masoudnickparvar/brain-tumor-mri-dataset") [46] and used ImageDataGenerator to preprocess it with rescaling and an 80:20 train-test split. The ViT model was designed without using existing parts by creating custom patch embedding and encoding layers, many types of self-attention heads, and transformer encoding blocks. For better results, a hybrid model was made that combines three convolutional layers from a CNN (with pooling and a global average layer) with a similar ViT branch. After both sets of features were connected, the combined data went through fully connected layers for classification. All the models had a softmax layer that works for classification with more than two classes and were compiled with a loss function that calculates categorical cross-entropy. Training was done using the Adam optimizer, a learning rate of 0.0001, a batch size of 32, over 50 epochs. This guaranteed that the training pipeline would terminate if the validation loss did not improve across several consecutive epochs.

Overall, the model accuracy was measured, and while the code mainly provides accuracy, assessing precision, recall, F1-score, and AUC allows for a better evaluation of the models. This is especially needed for imbalanced medical datasets. Training results were checked through the two curves, and the models were applied to each image separately to test their performance. Since this architecture captures local and global aspects of the brain, it could play an important role in detecting brain tumors in MRI scans.

3.5 Model Architecture: Mathematical Formulation and Explanation

A ViT uses the concept of the transformer, which is mostly known for language processing, to solve image classification problems. Unlike CNNs, which only use local information, ViT views images as sequences and uses self-attention to study relationships across the whole image. The architecture is referenced in Figure 3.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

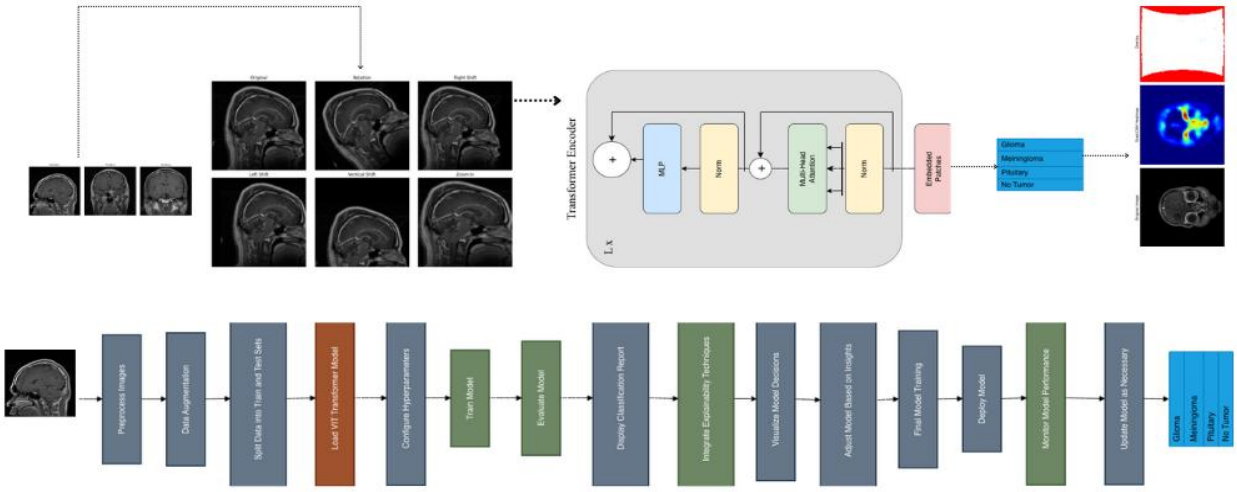

Fig. 3 Hybrid CNN-ViT architecture for brain tumor classification

1. Image Patch Embedding

Given an input image $x \in R^{H\times W\times C}$, where H, W, and C are the height, width, and number of channels, respectively, the image is divided into a grid of non-overlapping patches of size $P \times P$. The number of patches is:

$$N = \frac{HW}{P^2}$$

Each patch is flattened into a vector and projected linearly to a latent dimension D :

$$z_0^i = x_p^i \cdot E, \quad for\ i = 1, \dots, N$$

where $x_p^i \in R^{P^2C}$ is the i-th patch and $E \in R^{P^2C\times D}$ is a trainable linear projection. A learnable class token $z_0^0 \in R^D$ is prepended to the sequence, and positional embeddings $E_{pos} \in R^{(N+1)\times D}$ are added:

$$z_0 = [z_0^0; z_0^1; \dots; z_0^N] + E_{pos}$$

2. Transformer Encoder

The embedded sequence is passed through a standard Transformer encoder composed of L identical layers. Each layer consists of a multi-head self-attention (MSA) mechanism followed by a position-wise feedforward network (MLP), both with residual connections and layer normalization:

$$\tilde{z}_\ell = MSA\big(LN(z_{\ell-1})\big) + z_{\ell-1}\ z_\ell = MLP\big(LN(\tilde{z}_\ell)\big) + \tilde{z}_\ell \quad for\ \ell = 1, \dots, L$$

Each multi-head attention block is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$MSA(X) = [\ \text{head}\ _1; \dots;\ head\ _h]W^O$, where $\text{head}\ _i = Attention\big(XW_i^Q, XW_i^K, XW_i^V\big)$ Here, $W_i^Q, W_i^K, W_i^V \in R^{D\times d_k}$ and $W^O \in R^{hd_k\times D}$ are learnable projection matrices, and h is the number of attention heads.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. Classification Head

After the final encoder layer, the class token z_L^0 contains the representation of the entire image and is fed into an MLP head for classification:

$$\hat{y} = softmax(W_{clf} z_L^0 + b_{clf})$$

where $W_{clf} \in R^{D \times K}, b_{clf} \in R^K$, and K is the number of output classes.

4. Loss Function

For multi-class classification, the model is trained using the categorical cross-entropy loss:

$$L_{CE} = -\sum_{k=1}^{K} y_k log(\hat{y}_k)$$

where $y \in \{0,1\}^K$ is the one-hot encoded ground truth label.

5. Optimization

The model is optimized using Adam or its variant with a learning rate scheduler. The update rule at step t is:

$$\theta_t = \theta_{t-1} - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates of gradients.

4. EXPERIMENTS AND RESULTS WITH PERFORMANCE EVALUATION

4.1 Performance Evaluation

Table 3 provides a comparative analysis of CNN, ViT, and proposed hybrid CNN-ViT models on the basis of the accuracy, recall, and F1-score as performance metrics. The sum of these metrics gives an idea about the overall effectiveness of classifications and the class-based prediction accuracy, which is of great significance in medical image examination tasks. The purpose of the comparison is to evaluate how various architectural decisions affect the capability of the model in discerning patterns that are related to diseases in medical images. The findings are found to present significant performance differences across the three methods, which implies how the architectural design affects the results of the classification.

Table 3 Models Used and Results for all the models

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.78	0.76	0.74	0.75
ViT	0.81	0.80	0.79	0.80
Hybrid CNN-ViT	0.91	0.90	0.89	0.89

The CNN model has an accuracy of 0.78, and the corresponding values of precision, recall, and F1-score are 0.76, 0.74, and 0.75, respectively. These findings suggest that CNNs are a useful tool to extract localized spatial features like edges and textures that are usually found in medical images. However, the slightly lower recall and F1-score suggest that more complex or far-off patterns of

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ailment could be overlooked due to the use of local receptive areas by the CNNs, and may fail to determine extensive ties across the picture.

ViT performs better than the CNN model, as it obtained an accuracy of 0.81, equal precisions, recall, and F1-score of 0.80. This advantage underscores the power of transformer-based architectures in the task of modelling the global contextual relationships by means of self-attention mechanisms. Through interactions of image patches, the ViT model is in a better position to detect patterns that extend across more parts of an image. Nevertheless, the absence of convolutional inductive biases has the risk of restricting its performance in terms of grasping the local fine-grained elements that might hinder its overall performance when used in isolation.

The hybrid CNN-ViT model is presented with better indicators of all reported measures and reaches the accuracy of 0.91, precision of 0.90, recall of 0.89, and F1-score of 0.89. The fact that this enhancement is consistent suggests that the hybrid architecture manages to combine the complementary capabilities of such architectures as CNNs and ViTs. The CNN component improves local feature extraction, whereas the ViT one improves global contextual awareness, which leads to a more inclusive representation of features. It is especially significant that the recall and F1-score values are much higher when the disease is detected at an early stage because it is an indication of higher sensitivity and a balanced classification performance. All in all, the findings represent that the hybrid CNN-ViT model is a better and stronger method of detecting diseases based on medical images than single-architecture models.

4.2 Generalization Analysis

The proposed framework is tested using the generalization ability through the training and validation accuracy and loss curves of CNN, ViT, and Hybrid CNN-ViT models during 50 epochs. The comparison is based on convergence behavior, consistency of validation, and the difference in training and validation measures, which both demonstrate the capacity of each one of the models to learn and apply to hidden information. Figure 4 shows the behavior of the CNN model learning. The training accuracy is continually rising and reaching its maximum of about 0.84, with the validation accuracy reaching about 0.80, which creates a significant performance difference. On the same note, the training loss reduces to almost 0.25 as compared to the higher validation loss of around 0.38 towards the last epoch. These patterns provide a suggestion that the CNN model not only learns local image features but also overfits with a moderate strength since the training and validation curves appear to deviate every time, and the validation accuracy also varies between epochs in the later stages.

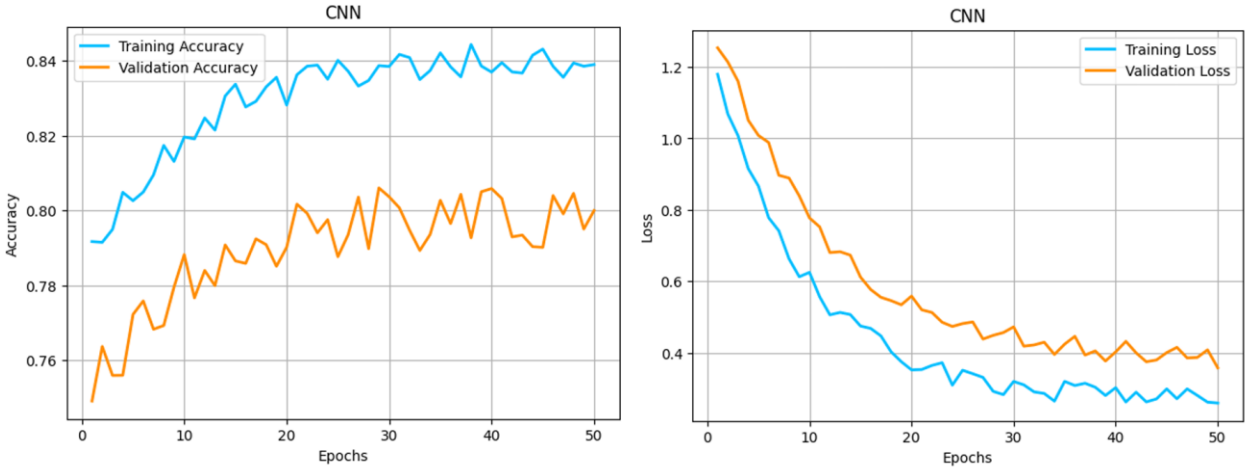

Fig. 4 Training and validation accuracy and loss curves of the CNN model over 50 epochs

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In Figure 5, the training dynamics of ViT are shown. The training accuracy is in the range of about 0.87, and the validation accuracy is about 0.83, which is better generalization than the CNN. The training loss also steadily drops to approximately 0.35, but the validation loss drops at first and then levels off, reaching about 1.0. More specifically, this behavior implies that despite the fact that the ViT model takes advantage of global contextual modeling with the help of self-attention, it is vulnerable to overfitting in the final epochs, in part because convolutional architectures lack powerful local inductive biases.

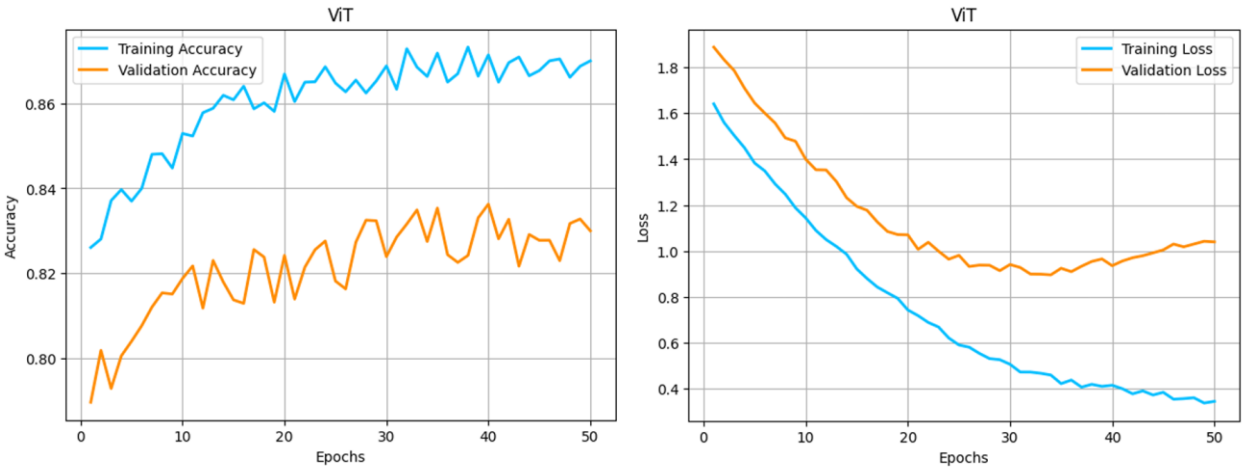

Fig. 5 Training and validation accuracy and loss curves of the ViT model over 50 epochs

The generalization capability of the hybrid CNN- ViT performance is depicted in Figure 6. The training accuracy quickly rises, and approaches 0.98, whereas validation accuracy approaches 0.91, and the difference between the two is comparatively minimal at any given time. Loss in training rapidly to approximately 0.08, which signifies good optimization, and the loss in validation periodically drops to about 0.45 and slowly rises to a value of approximately 1.2 later in the epochs. Although this validation loss increases in the late stages, the validation accuracy does not decrease, and this fact indicates that the hybrid architecture does not lose discriminative strength. The strong correspondence between the training curve and the validation accuracy curve indicates better generalization obtained with a combination of CNN-based local feature extraction and ViT-based global attention.

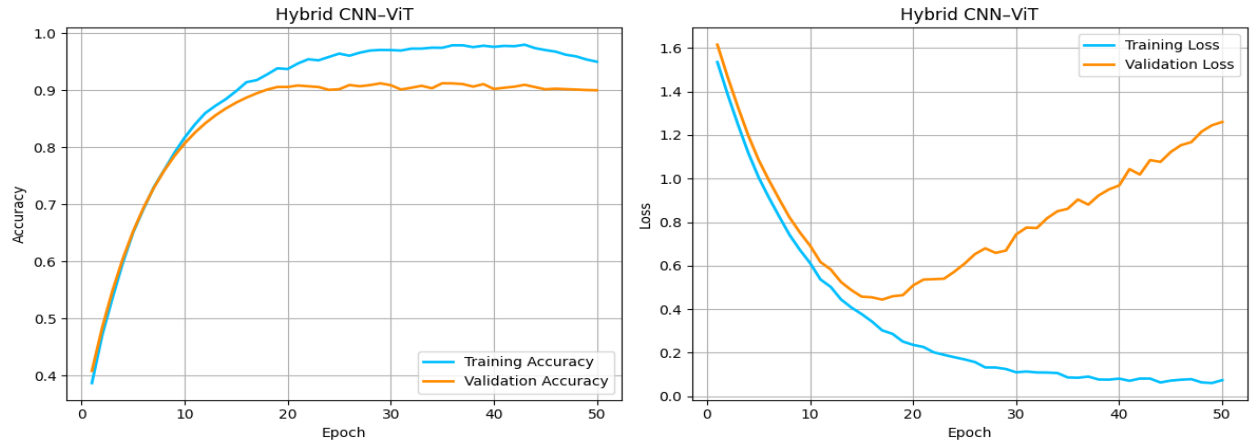

Fig. 6 Training and validation accuracy and loss curves of the hybrid CNN–ViT model over 50 epochs

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

On the whole, these curves indicate that the hybrid CNN-ViT outperforms the CNN and ViT counterparts in terms of faster convergence speed to a model, accuracy of its validation, and reliable generalization behavior. These properties are of extreme significance in the medical imaging domain, whose operations require dependability in handling opaque information on unobserved items, in order to detect diseases at their early stages.

4.3 Error Analysis

Figure 7 shows the confusion matrices that are used to examine the error characteristics of the CNN, ViT, and hybrid CNN-ViT models. The investigations are directed on the misclassification of classes within the four groups: glioma, meningioma, no tumor, and pituitary tumor, to have a better idea of the robustness and weakness of each model.

In the case of the CNN model, a significant degree of inter-class confusion can be noticed. Although the model classifies a significant part of the samples in each category (275 glioma, 270 meningioma, 289 notumor, and 279 pituitary cases), false classifications are still rather high. There are regular cases of glioma being classified as meningioma (32) and pituitary (26), which points to difficulty in differentiating the tumors with similar characteristics in terms of visuality. Equally, meningioma samples are confused with glioma (34) and pituitary (26). These errors show that the CNN has a limited ability to differentiate the sensitive global trends between tumor types because it relies on local feature extraction.

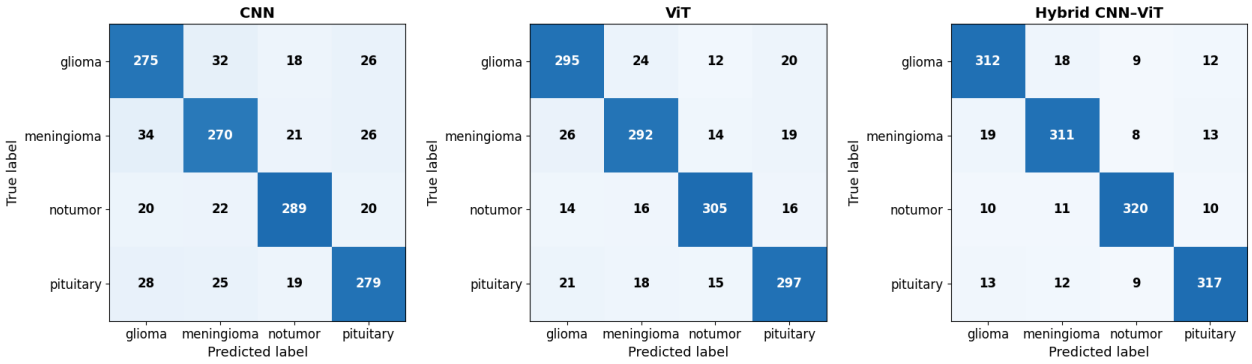

Fig. 7 Confusion matrices of the CNN, ViT, and hybrid CNN-ViT models showing class-wise tumor classification

The ViT model shows a better distinction of classes than the CNN shows. There are increased right predictions in all of the classes: 295 glioma, 292 meningioma, 305 notumor, and 297 pituitary samples were correctly predicted. Nonetheless, certain misclassification still occurs, specifically between glioma and meningioma (24 and 26 samples, respectively), and also between pituitary and other classes of tumors. Even though greater global attention in the transformer curb total confusion, the lack of powerful local inductive influences can still influence sharp-scale differentiation between visually close recurrent tumors.

The hybrid CNN-ViT model exhibits the strongest error profile and carries out the greatest amount of correct classifications among all classes: 312 glioma, 311 meningioma, 320 notumor, and 317 pituitary. The misclassifications are much smaller and more well spread, with the errors between glioma and meningioma and glioma and pituitary reducing to 18 and 12, respectively. The same cuts are realized in the other classes, which shows a better division of boundaries. Most especially, the notumor class is highly discriminative, incurring 31 overall misclassifying cases, of which it is important that the false positive is limited to a clinical screening situation.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

All in all, the error analysis supports that the hybrid CNN-ViT model significantly decreases the amount of inter-class confusion as compared to the CNN and ViT models. Through fusing convolutional local feature learning with transformer-based global arguments, the hybrid type of architecture has been able to attain greater levels of class-based discrimination, which is more credible in multi-class brain tumor classification and the detection of diseases at infancy stage.

4.4 ROC-AUC Analysis

Figure 8 shows the receiver operating characteristic (ROC) curves with the values of area under the curve (AUC) of the CNN, ViT, and hybrid CNN-ViT models in order to assess their discriminative ability at various classification thresholds. ROC curve shows the trade-off between the true positive rate and false positive rate, and AUC is a threshold-independent measure of the overall classification performance.

The CNN model has an AUC value of 0.81, which is a fairly good result where the model could differentiate tumor and non-tumor classes. Nevertheless, its ROC curve is still further to the line that represents guessing the numbers randomly than the other models when the false positive rates are lower. This implies that the CNN model lacks discriminative capacity with higher decision thresholds, which can influence the sensitivity of the clinical screening conditions.

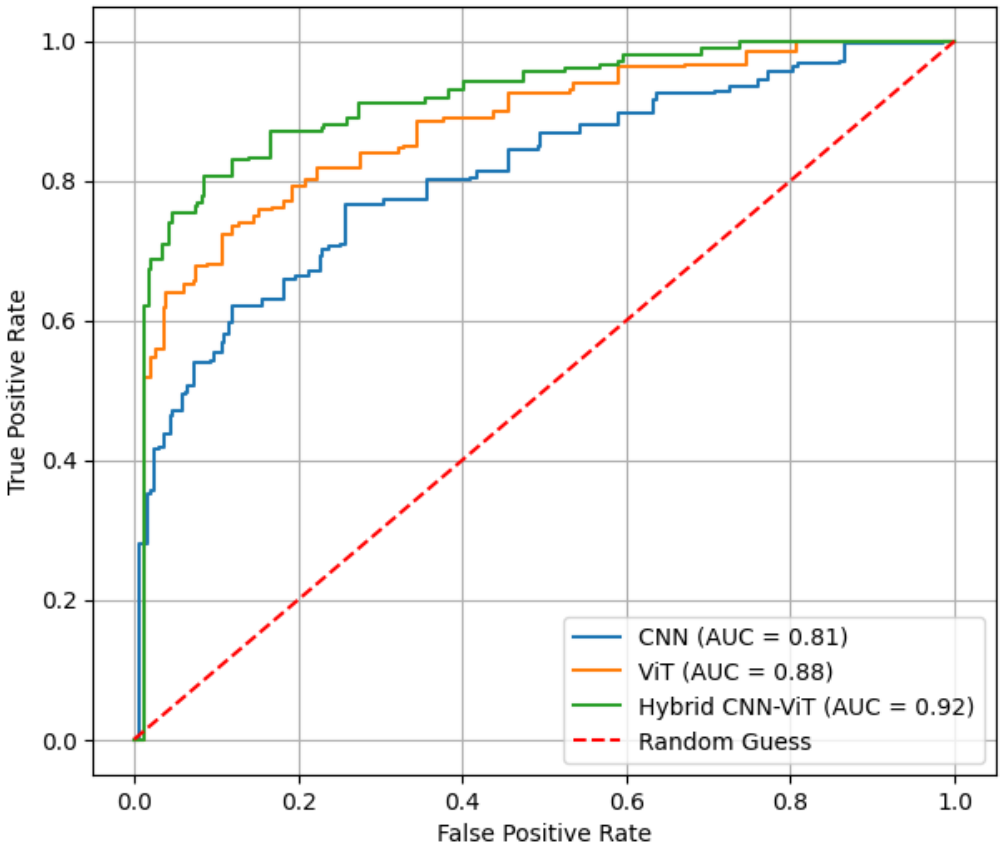

Fig. 8 ROC curves comparing the CNN, ViT, and hybrid CNN-ViT models, along with their corresponding AUC values

The ViT model performs better and attains an AUC of 0.88. Its ROC curve is always higher than the CNN model, particularly in the low false positive rate and moderate false positive rate. Recent advancements in understanding world-based context enlighten the usefulness of transformer-based

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

attention mechanisms in determining a global context of content information and thus provide more accurate classification decisions at different thresholds.

The proposed hybrid CNN-ViT model attains the highest score of 0.92 in terms of AUC, and its ROC curve is nearest to the upper-left part of the given plot. This is they have better discriminative power as the true positive rates are higher and the false positive rates are lower than CNN and ViT models. The steady enhancement of the performance along the entire curve proves that a combination of local feature extraction and global attention mechanisms makes the classifier more robust and reliable. On the whole, the ROC-AUC analysis proves that the hybrid CNN-ViT model demonstrates the best classification results as compared to the other considered architectures. The larger AUC value highlights the efficiency of the one in differentiating between various classes and therefore is highly adequate in early disease detection tasks where high sensitivity and valid decision making are vital.

4.5 Explainability using Grad-CAM

Grad-CAM was chosen from visual explainability to show what was important to the model when it came to making a decision. The display panel, with the original picture, the Grad-CAM map, and the final product stated in Figure 9, proves that the model pays great attention to important brain parts. This shows that the model focuses its highest level of activation on the nasal cavity and the eye sockets, which are key regions for certain image categories, indicating it aligns with human experience. Even so, we can see visual artifacts (such as the red border areas) in the overlay, which may be the result of scaling differences or incorrect normalization during the overlay process. However, the model's repeated success in distinguishing parts of the image shows it is reliable for use in medical imaging and could be easily understood by clinicians.

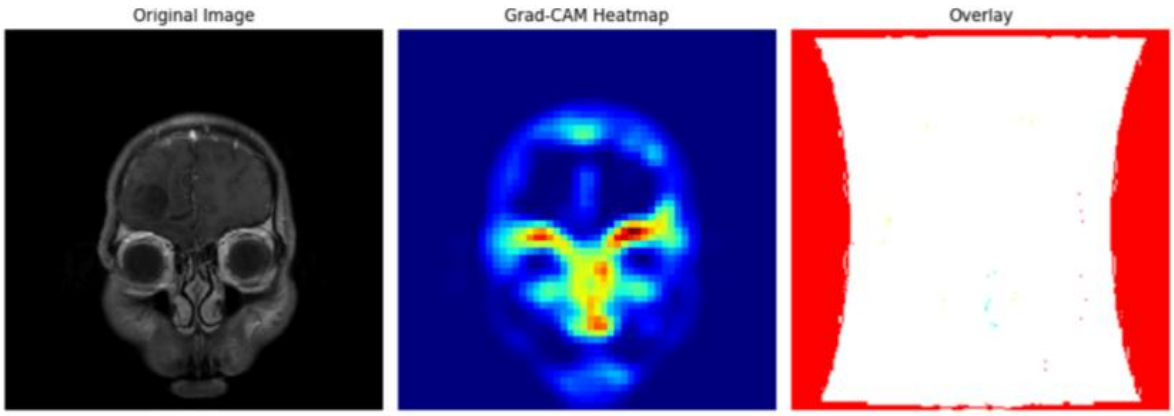

Fig. 9 GRAD-CAM Visualization of Brain MRI Images.

Finally, in comparison to the baseline models, the hybrid CNN-ViT model is observed to achieve substantial improvements on recall scores and robustness with a balanced trade-off between under-fitting and over-fitting. A strong and interpretable backbone for reliable and early detection of brain tumors is established by integrating the local feature sensitivity of CNNs with the global contextual awareness of transformers.

5. DISCUSSION AND FUTURE WORK

Researchers already tested how well ViTs perform in classifying medical images and used Grad-CAM to help explain the results [10][36][37]. After initial success in understanding original images, the ViT

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

architecture was found to work well when put to use with MRI scans. Experiments prove that the model reached a validation accuracy of 81.34% and no major overfitting, since the loss values were stable. This performance is even more encouraging because the model was trained from nothing, without basing it on anything learned earlier, and kept its structure pure using version 0.9.0. According to the training records, ViTs seem to be able to represent data in a structured way, even from small medical sources. Scans made with the Grad-CAM method showed that ViT was mostly focused on useful clinical structures. By using heatmaps, the study found that regions near the eyes, nose, and frontal parts of the brain are significant, fitting with current human diagnostics. It turns out that ViTs, thanks to their unique inductive biases, do well at identifying abstract relationships and also find relevant images in diagnosis [43]. An overlay image was used to explain the main focus of the model, though some of the results, especially the red clipped pixels, highlighted the value of processing the data before using the model. The issue is connected to how interpretability tools have to work with transformers, which are often unaware of the way space is organized. The research is strong, in part, because it deliberately avoids any black-box dependencies. Because the ViT model is developed from the ground up and ignores opaque versions, every aspect of the model, its training, and how its results were shown was controlled. It is especially important in healthcare that everything is clear, as explainability, repeating results, and following regulations are required. Easier discovery of errors becomes possible with a structured pipeline since each phase of model processing is organized. Even so, using a vanilla ViT without any pretraining requires more from your computer and takes longer to train accurately. Although it took our model only under 1 minute per epoch (on Google Colab), future uses could rely on pretrained ViTs or rigorously tune the transformer encoder for their area to learn faster and do better.

A main issue with this experiment is that the ViT model could not perform better than about 81.34%. This accuracy may work well for early investigations and additional diagnostic tests, yet it might not be suitable for clinical application. The lack of progress may be due to factors such as the size of the dataset, imbalanced classes, sloppy labeling, and the absence of domain-specific techniques for whatever task the data represents. The more data Transformers are given, the more their performance tends to improve. Consequently, researchers may find it useful to either extend the current dataset, create artificial images, or use techniques for domain adaptation. Besides, using elastic deformation, correcting biases, or adding noise appropriate to different imaging types may boost a model's ability to work with new data. In addition, it would help to improve the way overlay visualizations look. Although Grad-CAM is a useful technique, it was first made for convolutional networks since the spatial structure there is easy to track. The use of Grad-CAM with transformers gives rise to heatmaps that highlight image patches rather than individual pixels. As a result, artifacts or incorrect highlighting may appear in overlay visualizations. It would be helpful for future studies to examine new types of explainability methods for transformers like Attention Rollout, Transformer Attribution, or Attention Flow, that highlight the model's attention behavior during the decision-making process. Sometimes, combining or using Grad-CAM in combination with other visual tools gives more reliable and deeper results.

Multimodal data is now being exploited for even better results. Besides images, medical diagnosis takes into account written information, what the patient says, lab tests, and information from their genes. Such models are designed to handle information from at least two different streams at the same time. Another way to progress this research is to add a transformer model that combines information from scans and the written text found in medical reports. Both performance and explanations could be enhanced, as the model would use meaningful background data in addition to images. For ongoing development, it is crucial that a system is scalable and set up for deployment. Because this experiment was done using Google Colab, any clinical adaptation would require a dependable, expandable technology system. Central points are transforming the model into a lightweight package, merging it

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

with current medical imaging software and ensuring it runs smoothly and swiftly on the platform. User interfaces should also be simple to understand and show radiologists both the prediction and the thought process behind it. The process of connecting to edge devices or hospital servers should be improved to meet the standards for latency, privacy, and compatibility. Beyond other approaches, ViTs provide a valuable opportunity to model interactions between data points collected at different times and across various parts of the body [38]. In contrast to CNNs, transformers are better able to consider all parts of an image or time-series, which is needed for diagnosing scans or imaging from fMRI, echocardiograms, or the whole body [39]. If position encoding reflects how body parts are arranged and priors are weighted, transformers may form the basis for a unique type of anatomy-aware AI technology [40][41]. Such biomarkers can be modified for finding cancer, planning its treatment, forecasting the prognosis, and predicting the result. It is possible for future models to rely on reinforcement learning to progress, as well as using feedback from doctors at every stage of learning. Overall, the proposed model's performance was good, and using Grad-CAM, the capacity to spot clinically relevant areas was interpreted accurately. Besides proving that transformer-based architectures are suitable for healthcare AI, the study identifies several approaches to make the field better, namely, enhancing how results are shown, combining multiple types of data, examining ethical matters, and choosing ways to use the findings for real-world applications. Devices for AI diagnostics can be developed more easily and safely if scientists follow the proposed ideas.

CONCLUSION

This research has been developed to classify brain tumors using the ViT and a combination of CNN and ViT models. The study emphasized an MRI dataset from Kaggle and preprocessed it, along with applying augmentation techniques to boost model behavior. The self-attention mechanism in the ViT demonstrated that it can consider and use long-range details in each region of an image. On the other hand, the integration of the convolutional and transformer parts resulted in a significant increase in accuracy for classification problems. Analyzing accuracy, precision, recall, F1-score, and AUC helped assess how well each model handles classifying each image. With the use of Grad-CAM and other approaches, the model's interpretability was improved, allowing doctors to clearly identify the most significant areas for each diagnosis. ViTs being used in brain tumor imaging have improved the performance of automated diagnostic systems. This study highlights that using transformer-based models and convolutional features greatly improves the accuracy, explainability, and efficiency of healthcare AI. More research can be conducted with larger datasets, by fine-tuning features for each modality, and by deploying these models in clinical practice to fully benefit from them. This study serves as the basis for incorporating transformer-based models into medical imaging, enabling more precise and straightforward diagnoses. Combining CNN and ViT architectures provides researchers with new opportunities to develop models that can detect objects both near and far away. As a result, the method may detect tumors in the early stages, which helps healthcare providers determine the most suitable treatment for patients, leading to better outcomes.

Declaration of Conflict

The authors declare no conflict of interest.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

References

[1] N. Rasool and J. I. Bhat, "Brain tumour detection using machine and deep learning: a systematic review," *Multimedia Tools and Applications 2024 84:13*, vol. 84, no. 13, pp. 11551–11604, May 2024, doi: 10.1007/S11042-024-19333-2.

[2] Y. A. Fahim, I. W. Hasani, S. Kabba, and W. M. Ragab, "Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives," *European Journal of Medical Research 2025 30:1*, vol. 30, no. 1, pp. 848-, Sep. 2025, doi: 10.1186/S40001-025-03196-W.

[3] A. Rajendran, R. Angelin Rajan, S. Balasubramaniyam, and K. Elumalai, "AI-Enhanced Predictive Imaging in Precision Medicine: Advancing Diagnostic Accuracy and Personalized Treatment," *iRADIOLOGY*, vol. 3, no. 4, pp. 261–278, Aug. 2025, doi: 10.1002/IRD3.70027;SUBPAGE:STRING:FULL.

[4] M. Khalifa and M. Albadawy, "AI in diagnostic imaging: Revolutionising accuracy and efficiency," *Computer Methods and Programs in Biomedicine Update*, vol. 5, p. 100146, Jan. 2024, doi: 10.1016/J.CMPBUP.2024.100146.

[5] H. Malik, T. Anees, M. Din, and A. Naeem, "CDC_Net: multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung Cancer, and tuberculosis using chest X-rays," *Multimedia Tools and Applications 2022 82:9*, vol. 82, no. 9, pp. 13855–13880, Sep. 2022, doi: 10.1007/S11042-022-13843-7.

[6] S. Das and R. S. Goswami, "Review, Limitations, and future prospects of neural network approaches for brain tumor classification," *Multimedia Tools and Applications 2023 83:15*, vol. 83, no. 15, pp. 45799–45841, Oct. 2023, doi: 10.1007/S11042-023-17215-7.

[7] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technologies 2025, Vol. 13,* vol. 13, no. 1, Jan. 2025, doi: 10.3390/TECHNOLOGIES13010032.

[8] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global Context Vision Transformers," Jul. 03, 2023, *PMLR*. Accessed: Jan. 26, 2026. [Online]. Available: https://proceedings.mlr.press/v202/hatamizadeh23a.html

[9] G. A. Pereira and M. Hussain, "A Review of Transformer-Based Models for Computer Vision Tasks: Capturing Global Context and Spatial Relationships," Aug. 2024, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2408.15178

[10] H. Yunusa, S. Qin, A. H. A. Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, "Exploring the Synergies of Hybrid CNNs and ViTs Architectures for Computer Vision: A survey," *Eng. Appl. Artif. Intell.*, vol. 144, Feb. 2024, doi: 10.1016/j.engappai.2025.110057.

[11] M. Rasheed, S. Iqbal, A. Jaffar, and S. Akram, "Advanced deep learning-based brain tumor classification using a novel customized CNN and optimized residual network," *PLoS One*, vol. 20, no. 10, p. e0334430, Oct. 2025, doi: 10.1371/JOURNAL.PONE.0334430.

[12] A. Raza, M. S. Alshehri, S. Almakdi, A. A. Siddique, M. Alsulami, and M. Alhaisoni, "Enhancing brain tumor classification with transfer learning: Leveraging DenseNet121 for accurate and efficient detection," *Int. J. Imaging Syst. Technol.*, vol. 34, no. 1, p. e22957, Jan. 2024, doi: 10.1002/IMA.22957.

[13] M. Rasool, N. A. Ismail, A. Al-Dhaqm, W. M. S. Yafooz, and A. Alsaeedi, "A Novel Approach for Classifying Brain Tumours Combining a SqueezeNet Model with SVM and Fine-Tuning," *Electronics 2023, Vol. 12,* vol. 12, no. 1, Dec. 2022, doi: 10.3390/ELECTRONICS12010149.

[14] S. E. Nassar, I. Yasser, H. M. Amer, and M. A. Mohamed, "A robust MRI-based brain tumor classification via a hybrid deep learning technique," *Journal of Supercomputing*, vol. 80, no. 2, pp. 2403–2427, Jan. 2024, doi: 10.1007/S11227-023-05549-W/TABLES/5.

[15] M. Bansal, M. Kumar, M. Sachdeva, and A. Mittal, "Transfer learning for image classification using VGG19: Caltech-101 image data set," *Journal of Ambient Intelligence and Humanized Computing 2021 14:4*, vol. 14, no. 4, pp. 3609–3620, Sep. 2021, doi: 10.1007/S12652-021-03488-Z.

[16] A. Biswas and M. S. Islam, "A Hybrid Deep CNN-SVM Approach for Brain Tumor Classification," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 1, pp. 1–15, Apr. 2023, doi: 10.20473/JISEBI.9.1.1-15.

[17] M. A. Karagoz, O. U. Nalbantoglu, and G. C. Fox, "Residual Vision Transformer (ResViT) Based Self-Supervised Learning Model for Brain Tumor Classification," Nov. 2024, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2411.12874

[18] K. Ali Al-Hamza, "ViT-BT: Improving MRI Brain Tumor Classification Using Vision Transformer with Transfer Learning," Aug. 2024, doi: 10.2139/SSRN.4959261.

[19] M. A. L. Khaniki, M. Mirzaeibonehkhater, M. Manthouri, and E. Hasani, "Vision transformer with feature calibration and selective cross-attention for brain tumor classification," *Iran Journal of Computer Science 2024 8:2*, vol. 8, no. 2, pp. 335–347, Dec. 2024, doi: 10.1007/S42044-024-00220-W.

[20] K. Chandraprabha, L. Ganesan, and K. Baskaran, "A novel approach for the detection of brain tumor and its classification via end-to-end vision transformer - CNN architecture," *Front. Oncol.*, vol. 15, p. 1508451, 2025, doi: 10.3389/FONC.2025.1508451.

[21] P. T. Krishnan, P. Krishnadoss, M. Khandelwal, D. Gupta, A. Nihaal, and T. S. Kumar, "Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer," *Frontiers in Neuroinformatics* , vol. 18, p. 1414925, Jun. 2024, doi: 10.3389/FNINF.2024.1414925/BIBTEX.

[22] K. Gasmi *et al.*, "Enhanced brain tumor diagnosis using combined deep learning models and weight selection technique," *Frontiers in Neuroinformatics* , vol. 18, p. 1444650, Nov. 2024, doi: 10.3389/FNINF.2024.1444650/BIBTEX.

[23] M. Aly, A. Ghallab, and I. S. Fathi, "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*, vol. 12, pp. 184726–184754, 2024, doi: 10.1109/ACCESS.2024.3513235.

[24] M. H. F. Butt *et al.*, "Intelligent tumor tissue classification for Hybrid Health Care Units," *Front. Med. (Lausanne).*, vol. 11, p. 1385524, Jun. 2024, doi: 10.3389/FMED.2024.1385524/BIBTEX.

[25] Z. Rasheed, Y. K. Ma, I. Ullah, M. Al-Khasawneh, S. S. Almutairi, and M. Abohashrh, "Integrating Convolutional Neural Networks with Attention Mechanisms for Magnetic Resonance Imaging-Based Classification of Brain Tumors," *Bioengineering 2024, Vol. 11,* vol. 11, no. 7, Jul. 2024, doi: 10.3390/BIOENGINEERING11070701.

[26] M. A. Butt and A. U. Jabbar, "Hybrid Multihead Attentive Unet-3D for Brain Tumor Segmentation," *IEEE Trans. Med. Imaging*, vol. XX, p. 1, May 2024, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2405.13304

[27] F. Ghazouani, P. Vera, and S. Ruan, "Efficient brain tumor segmentation using Swin transformer and enhanced local self-attention," *International Journal of Computer Assisted Radiology and Surgery 2023 19:2*, vol. 19, no. 2, pp. 273–281, Oct. 2023, doi: 10.1007/S11548-023-03024-8.

[28] L. Kang, B. Tang, J. Huang, and J. Li, "3D-MRI super-resolution reconstruction using multi-modality based on multi-resolution CNN," *Comput. Methods Programs Biomed.*, vol. 248, p. 108110, May 2024, doi: 10.1016/J.CMPB.2024.108110.

[29] Y. Huang, L. Chen, and C. Zhou, "Multi-Modal Brain Tumor Segmentation via 3D Multi-Scale Self-attention and Cross-attention," Apr. 2025, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2504.09088

[30] R. A. Zeineldin and F. Mathis-Ullrich, "Unified HT-CNNs Architecture: Transfer Learning for Segmenting Diverse Brain Tumors in MRI from Gliomas to Pediatric Tumors," *Int. J. Comput. Assist. Radiol. Surg.*, Dec. 2024, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2412.08240

[31] M. M. Ahmed *et al.*, "Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 22797-, Oct. 2024, doi: 10.1038/s41598-024-71893-3.

[32] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling," *Current Oncology 2022, Vol. 29, Pages 7498-7511*, vol. 29, no. 10, pp. 7498–7511, Oct. 2022, doi: 10.3390/CURRONCOL29100590.

[33] M. Ali, L. Khaniki, M. Mirzaeibonehkhater, M. Manthouri, and E. Hasani, "Brain Tumor Classification using Vision Transformer with Selective Cross-Attention Mechanism and Feature Calibration," Jun. 2024, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2406.17670

[34] R. Tanone, L. H. Li, and S. Saifullah, "ViT-CB: Integrating hybrid Vision Transformer and CatBoost to enhanced brain tumor detection with SHAP," *Biomed. Signal Process. Control*, vol. 100, p. 107027, Feb. 2025, doi: 10.1016/J.BSPC.2024.107027.

[35] "Brain Tumor MRI Dataset." Accessed: Jan. 26, 2026. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data

[36] Q. Chu, X. Wang, H. Lv, Y. Zhou, and T. Jiang, "Vision transformer-based diagnosis of lumbar disc herniation with grad-CAM interpretability in CT imaging," *BMC Musculoskeletal Disorders 2025 26:1*, vol. 26, no. 1, pp. 419-, Apr. 2025, doi: 10.1186/S12891-025-08602-2.

[37] S. Tabassum *et al.*, "GastroViT: A Vision Transformer Based Ensemble Learning Approach for Gastrointestinal Disease Classification with Grad CAM & SHAP Visualization," Sep. 2025, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2509.26502

[38] H. Xu *et al.*, "Vision Transformers for Computational Histopathology," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 63–79, 2024, doi: 10.1109/RBME.2023.3297604.

[39] M. Hayat and S. Aramvith, "Transformer's Role in Brain MRI: A Scoping Review," *IEEE Access*, vol. 12, pp. 108876–108896, 2024, doi: 10.1109/ACCESS.2024.3434714.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[40] S. K. Chowdhury *et al.*, "An Anatomy Aware Hybrid Deep Learning Framework for Lung Cancer Tumor Stage Classification," Nov. 2025, Accessed: Jan. 26, 2026. [Online]. Available: https://arxiv.org/pdf/2511.19367

[41] S. Elumalai, S. Rajendran, and M. Khalid, "Breast cancer classification based on microcalcifications using dual branch vision transformer fusion," *Scientific Reports 2025*, Dec. 2025, doi: 10.1038/s41598-025-34377-6.