QUBAHAN JOURNAL OF MEDICAL SCIENCES
VOL. 1, Nov 2024
https://doi.org/10.48161/qjms.v1a17

HYBRID TRANSFORMER–CNN FRAMEWORK FOR EXPLAINABLE MULTIMODAL BRAIN TUMOR CLASSIFICATION USING MULTI-SEQUENCE MRI

Kafui Tsoeke Agbevanu[1], Mohamed Merza Khalil Ebrahim Al.Rayes[2*]

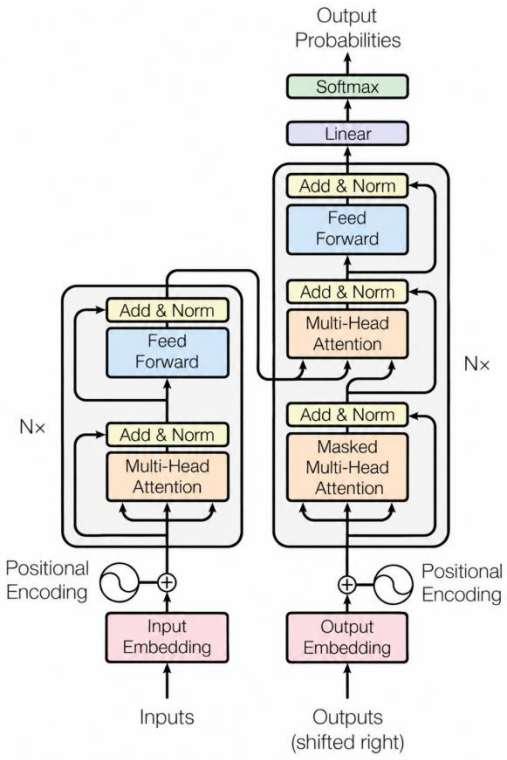
[1] Computer Science, Ho Technical University, HO, Ghana;
[2] Physiotherapy, College of Health Sciences, University of Bahrain, Ahlia University, Manama, Bahrain;
***Corresponding author:** denesh.sooriamoorthy@apu.edu.my**.**

ABSTRACT: There has been significant progress in the field of medical image analysis due to advancements in artificial intelligence (AI) and deep learning technologies, enabling automated, more accurate, and more efficient disease diagnosis. An AI-based model, Hybrid CNN-Transformer, is introduced for explainable brain tumor classification using MRI brain images. The proposed framework leverages the feature-extraction capabilities of CNNs and the context-learning capabilities of Transformers to achieve better diagnostic outcomes and increased interpretability. For this purpose, a freely available MRI brain tumor dataset, comprising glioma, meningioma, pituitary, and no-tumor categories, was used to develop and validate the proposed framework. Prior to developing the model, preprocessing steps, such as resizing, normalization, and data augmentation, were applied to MRI images to enhance image quality and improve the model's generalization capacity. The DenseNet121 CNN was used to extract features in the spatial domain, whereas the Transformer encoder was used to capture contextual relationships via self-attention mechanisms. Experimental evaluation has been conducted using accuracy, precision, recall, F1 score, specificity, the confusion matrix, and ROC-AUC. Comparison with CNN and Transformer models shows that the proposed approach achieves competitive diagnostic accuracy while offering better explainability. The results clearly indicate considerable promise in using hybrid architectures for AI-based radiological diagnostics.

Keywords: Grad-CAM, MRI, explainable artificial intelligence (XAI), ReLU.

I. INTRODUCTION

Medical image analysis has been revolutionized by artificial intelligence (AI) and deep learning, enabling accurate, automated diagnosis of diseases from complex imaging data. In clinical practice, medical imaging techniques such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasound are routinely used to identify abnormalities, track disease progression, and plan treatment (Alruwaili et al., 2024). Routine diagnostic testing usually depends on the radiologist's work, which can lead to delays, inconsistent interpretation, and misdiagnosis. This has led to the development of systems powered by artificial intelligence (AI) to standardize and enhance diagnostic efficiency and reliability. This has attracted considerable interest in developing AI-powered systems to standardize and enhance diagnostic efficiency and reliability. Medical images provide essential information about the spatial characteristics of medical entities, which has become crucial for classification and detection tasks (Obuchowicz et al., 2024).

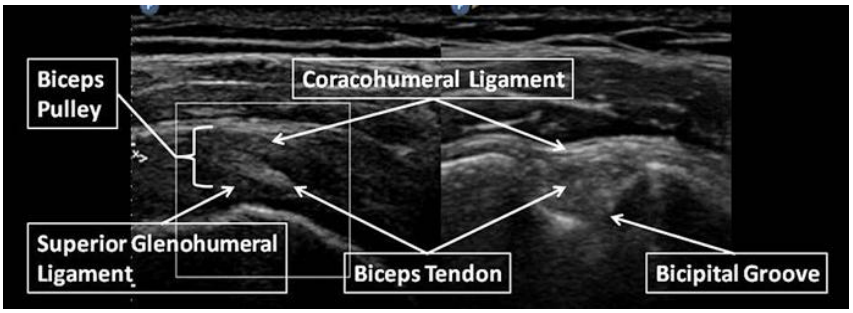


(Source: Zhang et al., 2024).

FIGURE 1. The Transformer structure, consisting of an encoder and a decoder.

The capabilities of deep learning, especially Convolutional Neural Networks (CNNs), have proven to be of great value in the classification and detection of medical images. Most recently, Transformer-based models have become powerful, leveraging self-attention mechanisms to learn long-range dependencies and global contextual information (Zhang et al., 2024). These advances have enhanced the ability of automated diagnostic systems to detect finer patterns in medical images. Moreover, with the advent of multimodal medical imaging, information from various modalities is complementary and thus increasingly important. Incorporating data from MRI, CT, and ultrasound can, therefore, improve diagnostic accuracy and aid holistic clinical decision-making in the modern healthcare setting.

Although tremendous advances have been made in the field of medical imaging diagnosis, there are still some gaps in current medical imaging models. Very few existing diagnostic systems are developed for the analysis of imaging data from a single modality, such as MRI-only or CT-only systems, thereby limiting the use of complementary information across modalities. The use of MRI, CT, and ultrasound images will provide different views of anatomical and pathological details, and a single imaging modality may not provide sufficient diagnostic information or understanding of the disease (Toh, 2024). However, traditional deep learning models, such as CNN-based models, primarily rely on extracting local features from images and may not adequately model long-range context in medical images. Transformer models have achieved good performance in global feature learning, but to balance spatial and contextual representation learning, it is necessary to integrate them with CNN architectures.



(Source: Toh, 2024).

FIGURE 2. Anatomy of the shoulder joint seen on ultrasound.

Another challenge is that many AI diagnostic systems lack interpretability. However, because the process of utilizing AI for decision-making is opaque, it is difficult to demonstrate that medical professionals can trust black-box AI models (Erben, et al., 2024). Additionally, current learning approaches often lack efficient feature fusion and explainability. All these difficulties emphasize the importance of developing a diagnostic strategy that is integrated, interpretable, and able to enhance classification accuracy and clinical reliability.

Previous research in medical diagnosis has mainly focused on CNN-based and Transformer-based Models. Few studies have integrated both models into one. While CNNs can effectively learn local spatial information and Transformers can learn global context, there are still few studies that integrate the complementary strengths of both approaches for multimodal medical image analysis. Moreover, most existing diagnostic systems are based on a single imaging modality, and few studies have focused on using multiple modalities, such as simultaneous MR, CT, and US, for a more comprehensive diagnosis of disease. Another notable drawback is the lack of use of explainable AI (XAI) methods in multimodal diagnostic processes. The majority of deep learning models in place are 'black-box ', which makes them less transparent and less trusted by healthcare practitioners. Thus, a consistent, interpretable multimodal feature fusion and clinically reliable diagnostic result-delivering EH OR Enhanced Hybrid (Transformer-CNN) framework is still needed.

The study aims to develop an explainable hybrid Transformer-CNN framework that integrates MRI, CT and ultrasound images for accurate and interpretable multimodal medical image diagnosis. The objectives of the research are to preprocess and integrate MRI, CT, and ultrasound datasets for multimodal medical image analysis. To develop a hybrid CNN-Transformer architecture for extracting local and global features from medical images. To implement explainable AI techniques for improving interpretability and clinical trustworthiness. To evaluate the proposed framework using standard medical imaging performance metrics and comparative analysis.

The research question is: How can a hybrid Transformer-CNN architecture improve multimodal medical image diagnosis accuracy? How effective is multimodal fusion using MRI, CT, and ultrasound data in disease classification? Can explainable AI techniques improve the interpretability and reliability of automated diagnosis systems? How does the proposed framework perform compared to conventional CNN-based and Transformer-based models?

The proposed study is important as it helps develop accurate, interpretable, and clinically reliable AI-assisted diagnostic systems for modern healthcare applications. The study combines MRI, CT, and ultrasound imaging modalities to learn multimodal features to enhance disease diagnosis. Multiple imaging modalities enable the framework to acquire complementary anatomical and pathological data, potentially increasing diagnostic accuracy compared with single-modality systems (Kumar, 2022). Moreover, the study advances the development of hybrid deep learning architectures that integrate CNNs and Transformers into a single framework. CNN models work well for extracting local features, and combined Transformer networks work well for capturing global features using self-attention mechanisms (Niu et al., 2023). The combined use of these techniques should improve overall classification performance and robustness.

In addition, the use of explainable AI techniques such as Grad-CAM, SHAP, and attention visualization enhances the interpretability and transparency of automated diagnosis systems (Nazir, 2024). This can help build trust in the clinical setting and help health care workers understand AI-generated predictions. Results

from the current study can thus inform future developments in intelligent healthcare systems, precision medicine, and computer-aided medical diagnosis technologies.

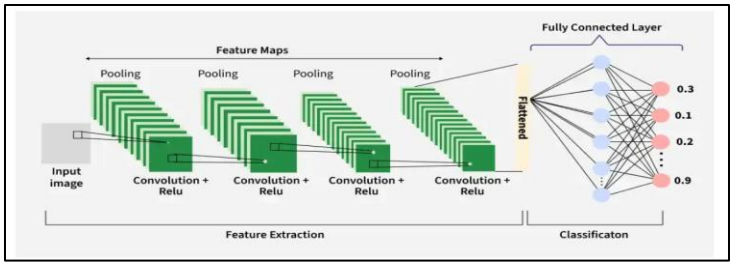
II. LITERATURE REVIEW

1. DEEP LEARNING IN MEDICAL IMAGING

Deep learning has emerged as one of the most significant technologies in medical image analysis and disease diagnosis. It employs artificial neural networks to automatically extract useful patterns from medical images without the need for manual feature extraction. Medical imaging, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-ray, and ultrasound, generates large volumes of image data that deep learning models can effectively process. These models are designed to assist medical workers in detecting disease, classifying abnormalities, segmenting organs, and supporting decision-making (Zhou et al., 2021). Deep learning is effective in medical imaging for diagnostic accuracy and efficiency. AI-powered systems achieve high precision in detecting tumors, lesions, fractures, and other pathological changes, and they reduce radiologists' workload. In automated image analysis, deep learning has been effectively used to diagnose cancers, cardiovascular diseases, neurological disorders, retinal diseases, and respiratory conditions, among others (Khalifa, 2024). Image classification, object detection, and image segmentation are popular tasks that are processed with deep learning models. State-of-the-art architectures have also improved the ability to learn effective visual patterns in medical images. These include Convolutional Neural Networks (CNNs), U-Net models, and recently released Transformer-based networks. Though there are developments, issues such as limited annotated datasets, model interpretability, and fairness remain considerations in clinical applications. The accuracy, reliability, and clinical acceptance of AI-aided diagnostic systems are likely to increase with continued advances in deep learning algorithms in the future of healthcare.

2. CNN-BASED MEDICAL DIAGNOSIS MODELS

Convolutional Neural Networks (CNNs) are one of the deep learning models most commonly used in medical image diagnosis. Medical images can also be used to automatically learn and retrieve important image features, such as edges, textures, shapes, and disease patterns, using CNNs. Several CNN architectures, including ResNet, DenseNet, and EfficientNet, have been successfully applied to brain tumor analysis. These models have demonstrated strong performance in distinguishing between different tumor types and improving classification accuracy. CNNs mainly focus on local image features and may not effectively capture long-range relationships across the entire MRI image. ResNet (Residual Network) is one of the most powerful CNN structures,ures as it adds connections that enable information to flow through deep layers without losing significant features. This architecture helps mitigate the vanishing gradient problem and enables the training of extremely deep networks. ResNet has been popular for detecting lung diseases, brain tumors, and breast cancer in medical images due to its strong classification performance. Another important architecture is DenseNet (Densely Connected Network), which links all layers to each other, promotes feature sharing, and enhances information flow across the network (Hou et al., 2024). This architecture minimizes the number of parameters while achieving high accuracy. DenseNet has already shown strong performance in medical image classification, segmentation, and object detection, particularly in cancer diagnosis and radiological image analysis. EfficientNet is a more recent CNN architecture optimized to achieve high accuracy with reduced computational cost. It employs a balanced scaling algorithm that trades off network depth, network width, and image resolution. EfficientNet has demonstrated strong performance in skin disease diagnosis, chest X-ray analysis, and other diagnostic tasks, while also requiring fewer computational resources than most conventional CNN models.

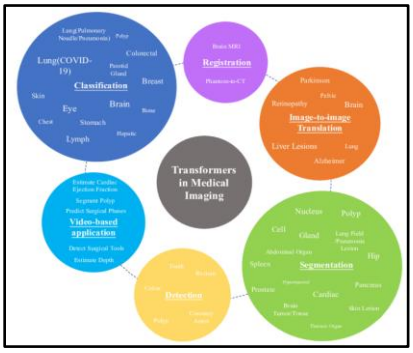


(Source: GeeksforGeeks, 2017).

FIGURE 3. Convolutional neural networks.

3. TRANSFORMER MODELS IN HEALTHCARE IMAGING

Transformer models have recently become an important advancement in medical image analysis. Transformers include a self-attention mechanism that allows them to capture local and global relationships in an image, in contrast to classical Convolutional Neural Networks (CNNs), which primarily capture local image features. This capability enables Transformer models to capture long-range dependencies and complex patterns, making them useful in healthcare imaging applications (Henry, 2022). The Vision Transformer (ViT) is one of the most popular Transformer-based architectures. Vision Transformers break images into patches and compare relationships between patches in order to determine significant diagnostic patterns. Transformer-based models have demonstrated strong potential for brain tumor classification by enhancing feature representation and classification performance. They are very useful in more complex medical imaging because they can model intricate spatial relationships. The latter is an added benefit of ViTs, as they capture global contextual features compared to traditional CNNs, which is especially useful for detecting disease patterns distributed across different areas of an image. Transformer models have been shown to be effective in detecting brain tumors, lung diseases, retinal diseases, and breast cancer using medical images. Self-attention processes enhance feature representation by assigning greater significance to clinically important areas of images and diminishing the impact of irrelevant background data (Xia, 2023). This increases the tendency of the model to give accurate diagnostic predictions.



(Source: ResearchGate).

FIGURE 4. Applications of transformers in medical image analysis.

4. MULTIMODAL MEDICAL IMAGE FUSION

Multimodal medical image fusion is a significant approach for merging data from various imaging modalities to enhance disease diagnosis and clinical decision-making. There are various medical imaging practices which complement each other in the human body. Magnetic Resonance Imaging (MRI) can provide detailed soft-tissue imaging, Computed Tomography assists in providing clear anatomical structures, and ultrasound imaging provides real-time images of organs and tissues. Integrating these modalities provides a more detailed picture of disease conditions than using a single imaging source (Anghel et al., 2024). Multimodal learning is concerned with the extraction and integration of useful features from various types of medical information. There are fusion methods that aim to create a single representation to extract

information from multiple imaging modalities. Brain tumor diagnosis often requires information from multiple MRI sequences because each sequence highlights different tissue characteristics. The most common MRI sequences are T1-weighted, T1-contrast-enhanced (T1ce), T2-weighted, and Fluid-Attenuated Inversion Recovery (FLAIR) images (Zhao et al., 2020). Integration of information from these sequences can provide a more comprehensive picture of tumor structures and surrounding tissues. Multi-sequence MRI fusion methods are designed to combine complementary information across multiple sequences to enhance classification results. Deep learning-based fusion methods have become increasingly popular because they can automatically learn relationships between multiple MRI modalities. Additionally, attention-based fusion mechanisms can automatically select the most informative features from each modality and enhance diagnostic performance.

5. EXPLAINABLE AI IN MEDICAL DIAGNOSIS

Explainable AI has gained significant importance in medical diagnosis due to the properties of many deep learning models. Openness and trust are essential practices, especially when AI systems are used to assist in making important diagnostic decisions. Explainable AI techniques help explain the rationale behind model forecasts, enhancing trustworthiness, accountability, and acceptance among health care providers (Tjoa, 2020). Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most popular explainability methods. Grad-CAM produces a heatmap showing the regions of a medical image that contribute most to a model's prediction. Using Grad-CAM, it is possible to apply it to tumor detection and disease classification, allowing clinicians to see with their own eyes that the model focuses on significant anatomical structures and pathological features (Nazir, 2024). Another important explainability method is SHapley Additive Explanations (SHAP), which helps researchers and clinicians understand the effects of various image characteristics on diagnostic outcomes. SHAP explains both at the global and local levels, which is helpful for estimating the model's behavioral patterns in general and for making predictions at the individual patient level. Another explainable method is attention visualization, which shows attention maps indicating which parts of the image receive the most attention during prediction. This assists in understanding medical image processing through self-attention mechanisms and in finding clinically meaningful patterns. Grad-CAM, SHAP, and attention visualization are crucial to the interpretability of current AI-based medical diagnosis systems.

6. LITERATURE GAP SUMMARY

The latest developments in deep learning have greatly enhanced medical image diagnosis through Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), multimodal learning, and explainable artificial intelligence (XAI). CNN-based models such as ResNet, DenseNet, and EfficientNet have shown good performance in local image feature extraction and disease pattern identification across MRI, CT, X-ray, and ultrasound images. Transformer-based models have demonstrated encouraging performance on long-range dependencies and global contextual information, leveraging self-attention mechanisms that enhance diagnostic performance across several healthcare imaging tasks. Although these developments exist, several gaps persist in the research. Most extant research investigates either CNN-based or Transformer-based architectures, but not both simultaneously, instead leveraging the unique benefits of each within a single framework. CNNs are efficient at learning local spatial characteristics, whereas Transformers are efficient at learning global contextual relationships. The other gap is the lack of multimodal imaging data, as the available diagnostic systems rely predominantly on a single imaging modality, which may be insufficient to provide detailed information on complex diseases (Tjoa, 2020). The combination of MRI, CT, and ultrasound data has not been sufficiently studied, even though such integration offers potential advantages. Explainable AI methods such as Grad-CAM, SHAP, attention visualization, and other techniques have gained attention but are not always incorporated into multimodal diagnostic systems.

III. METHODOLOGY

1. FRAMEWORK ARCHITECTURE

The proposed framework utilizes a hybrid CNN-Transformer architecture for automated brain tumor classification from MRI images. The framework combines the local feature extraction capability of CNNs with the global contextual learning ability of Transformer models to improve diagnostic performance. MRI images are preprocessed through resizing, normalization, and augmentation techniques to improve image quality. The preprocessed images are then passed to a DenseNet121-based CNN backbone (Poudyal et al., 2024). The extracted feature maps are subsequently forwarded to the Transformer encoder, and self-attention mechanisms learn long-range contextual relationships. The Transformer output, combined with the representations generated by the CNN, creates a complete feature embedding that captures local pathological features. Grad-CAM explainability techniques are integrated into the framework to visualize diagnostically significant regions and enhance the interpretability of model predictions (Guluwadi, 2024).

2. DATASET COLLECTION

For the experiment, the Kaggle MRI dataset of brain tumors was used to train and validate the proposed hybrid architecture. This data consists of MRI images belonging to the following four classes: glioma tumor, meningioma tumor, pituitary tumor, and no tumor (Kaggle.com, 2024). This particular dataset was selected for its balanced and relevant collection of MRIs, which is useful for automated brain tumor classification and medical image analysis. The MRI image dataset includes two directories containing JPEG images: the training and testing datasets. The former data directory was used to train the proposed model and optimize its parameters, while the latter contained the data used to validate the proposed framework. The dataset contains sufficient variation in pathology shape, intensity, texture, and anatomical structure. MRI was chosen as a tool of interest for this research because it is widely considered one of the best options for brain tumor diagnosis.

3. DATA PREPROCESSING

Data processing was performed to improve image quality, enhance the model, and ensure consistency during training and evaluation. The images produced from the MRI data differed in image size, orientation, intensity, and noise levels. Prior to feeding the images into the proposed Hybrid CNN-Transformer, a few preprocessing methods were utilized. Each MRI image was resized to 224 x 224 pixels to ensure that the input size was always the same (224 x 224) to fit the DenseNet121 backbone architecture. Imagnormalizationon was then applied tstandardizese pixel intensity distributions anstabilizese model convergence during training. The normalization process was performed using the following equation.

$$I_{norm} = \frac{I - \mu}{\sigma}$$

Where I_{norm} is the normalised image, I is the original MRI image, μ is the mean pixel intensity, and σ denotes the standard deviation. The augmentation operations included random horizontal flipping, rotation, and minor spatial transformations. These methods increased the diversity of the datasets and enabled the framework to learn invariant tumor representations across varying imaging conditions. The PyTorch transformation pipelines normalized the images to tensor format. The processed MRI images were then partitioned into training and testing datasets to train the model and evaluate its performance. This step ensured that the MRI data was normalized, denoised, and optimized to learn features efficiently using Transformers.

4. CNN FEATURE EXTRACTION MODULE

The CNN feature extraction module was developed to learn discriminating spatial features from MRI brain tumor images. DenseNet121 was employed as the main convolutional backbone due to its excellent feature propagation and effective gradient flow. DenseNet architectures are widely applicable to medical imaging because they can extract fine-grained pathological details from complex image data. The preprocessed MRI images were fed into the CNN module, where multiple convolutional operations were performed to extract hierarchical image representations. The convolutional layers learned low- and high-

level features, such as tumor boundaries, edges, and structural abnormalities. The convolutional operation can be represented as follows.

$$F(i,j) = \sum_m \sum_n I(i-m, j-n) \cdot K(m,n)$$

Where F(i, j) denotes the generated feature map, I denotes the input MRI image, and K denotes the convolution kernel. Consecutive convolutional operations were followed by the addition of the Rectified Linear Unit (ReLU) and Batch Normalization activation functions to enhance training stability (Al-Shoukry & Musa, 2024). The spatial dimensionality was then reduced using max-pooling layers, while important tumor features were retained. The DenseNet121 backbone produced high-dimensional feature representations that captured clinically relevant tumor features. The spatial dimensionality was then reduced using max-pooling layers without losing significant tumor characteristics (Mohamed et al., 2022).

5. TRANSFORMER FEATURE LEARNING MODULE

A transformer feature-learning component was added to the proposed model to capture global contextual relationships in MRIs of brain tumors. Although CNNs excel at learning localized features, they cannot model dependencies that span longer distances across different parts of an image. Consequently, a Transformer encoder was added to enhance context awareness using self-attention. The CNN component generated high-dimensional feature vectors, which were fed to the Transformer encoder to learn global features. Using the self-attention mechanism, the Transformer architecture was able to capture relationships among different feature regions in MRI images (Krishnan et al., 2024). The self-attention mechanism used in the Transformer module can be mathematically expressed as:

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

Where Q is the query matrix, K is the key matrix, V is the value matrix, and dk denotes the dimensionality scaling factor. The attention mechanism computed similarities among feature representations and assigned greater weight to medically important regions in the images. Multi-head attention mechanisms were implemented in the model to learn multiple contexts simultaneously.

3.6 Multimodal Fusion Strategy

The multimodal fusion strategy was designed to integrate CNN-generated spatial features with Transformer-based contextual representations (Guo et al., 2022). The CNN module extracted localized pathological features, such as tumor boundaries and texture patterns. These feature representations were fused to generate a unified embedding for accurate classification. The fusion process can be mathematically represented as:

$F_{fusion} = \alpha F_{CNN} + \beta F_{Transformer}$

Where F_{CNN} is the CNN feature vectors, Transformer denotes the transformer contextual features, and α and β are the weighting coefficients.

6. EXPLAINABILITY MODULE

A module for explainable artificial intelligence (XAI) was incorporated into the developed framework to enhance interpretability and clinical validity (Qian et al., 2023). Grad-CAM was used to visualize the image regions that contributed most to the classification. The XAI module provided heat maps indicating tumour segments with high diagnostic relevance in the MRI images. The greater the activation intensities, the more highly the segments were prioritized by the algorithm in prediction. This facilitated visualizing the localized tumor structures and the surrounding context that the CNN and Transformer modules learn.

7. CLASSIFICATION LAYER

The classification layer was responsible for generating the final tumor predictions using the fused CNN and Transformer feature representations. Fully connected dense layers were employed to map the extracted feature integrating 4 output categories. To reduce overfitting and improve model generalization, dropout

regularisation was added. A softmax activation function was used to produce the final prediction probabilities.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$$

Where P(yi) represents the probability of class i, zi denotes output logits, and N is the total number of classes. The class with the highest probability was selected as the final diagnostic prediction.

IV. EXPERIMENTAL SETUP

1. SOFTWARE TOOLS

The Hybrid CNN–Transformer was coded in Python with various open-source deep learning and image processing libraries. To be compatible with various applications in artificial intelligence, machine learning, and medical image analysis, Python was chosen as the software tool. The main deep learning library used in this study was PyTorch, which provided versatility for building, training, and evaluating models. PyTorch was used for implementing the CNN-only, Transformer-only, and Hybrid CNN–Transformer architectures and for efficient computation on a GPU. Image loading, processing, augmentation, and obtaining DenseNet121 models were carried out using the torchvision library. Operations such as resizing, normalization, and data transformation were performed on the images using torchvision transforms. Throughout the experimental process, OpenCV was used for image handling and visualization, while NumPy was used for numerical computations and array manipulations. Scikit-learn was used for performance evaluation and statistical analysis. The measures of accuracy, precision, recall, F1 score, confusion matrices, and ROC-AUC were computed using this library. The data were organized, and the results were managed using Pandas. Moreover, graphical representations such as confusion matrices and ROC curves were produced using the Matplotlib and Seaborn libraries. Using the Grad-CAM library, explainable AI analysis was performed to visualize the regions in an image that were most important in the model's predictions. All experiments were conducted in the Google Colab environment, a convenient cloud-based platform for training and evaluating models (Erben et al., 2024).

2. HARDWARE CONFIGURATION

The experimental implementation was performed in the Google Colab environment, which includes a GPU. Google Colab was chosen because it provides access to high-performance computational resources without relying on local hardware to run the program. The platform enabled faster model training and reduced the computational load associated with deep learning applications (Erben et al., 2024). Experiments were carried out with the help of the NVIDIA Tesla GPUs provided by Google Colab. For the convolutional neural network and Transformer architectures, training was significantly faster on GPU compared to CPU-only training. Furthermore, the runtime system had sufficient RAM and storage capacity to load the MRI image datasets, preprocess the images, train the models, and generate the evaluation results. The software environment with all the needed deep learning libraries was configured using Python 3.x. PyTorch was used to implement the model with GPU (CUDA) support. This hardware setup ensured efficient execution of the CNN-only, Transformer-only, and Hybrid CNN–Transformer models and helped ensure reproducible experimental results for the study.

3. HYPERPARAMETER SETTINGS

The hyperparameters were set for optimum training and classification. Before giving the MRI images to the neural network architectures, they were resized to 224 × 224 pixels. To prevent overfitting and enhance the model's generalizability, the data were augmented using techniques such as random Horizontal Flipping and random Rotation. Also, to avoid overfitting, 0.3 dropout was added to the classification layers to improve classification robustness. Since it is a multi-class classification problem, the objective function used was Cross-entropy loss. These hyperparameter configurations enabled the researcher to train all architectures under evaluation and made it easy to compare the performance of the CNN-only, Transformer-only, and Hybrid CNN–Transformer models (Liao et al., 2022).

V. RESULTS AND DISCUSSION

1. PERFORMANCE EVALUATION

MRI brain tumor images across four classes (glioma, meningioma, pituitary tumor, no tumor) were used to evaluate the proposed Hybrid CNN-Transformer framework. The experiments were conducted on Google Colab with a GPU. The performance of the framework was evaluated using widely used medical image classification metrics, including accuracy, precision, recall, F1-score, specificity, and ROC-AUC. The experimental results showed that the CNN-only architecture outperformed the other architectures evaluated. The CNN model obtained an accuracy of 80.46%, precision of 85.78%, recall of 80.46%, and F1-score of 77.20%. The proposed Hybrid CNN-Transformer framework achieved an accuracy of 74.37%, precision of 82.41%, recall of 74.37%, and an F1-score of 70.69%. The Transformer-only model, on the other hand, achieved significantly lower performance, with an accuracy of 24.11%, precision of 24.09%, recall of 24.11%, and an F1-score of 20.45%.

Table 1. Performance evaluation results.

Model	Accuracy	Precision	Recall	F1-Score
CNN-only	80.46%	85.78%	80.46%	77.20%
Transformer-only	24.11%	24.09%	24.11%	20.45%
Hybrid CNN-Transformer	74.37%	82.41%	74.37%	70.69%

The analysis of the confusion matrix allowed us to compare the CNN-only, Transformer-only, and Hybrid CNN-Transformer frameworks. The CNN-only model achieved good classification rates across most MRI tumor categories, especially Classes 1 and 2, where a large number of samples were correctly classified. The Transformer-only model did not perform as well because many MRI sample classifications were misclassified due to its limited ability to extract local features. The Hybrid CNN-Transformer framework achieved better-balanced classification rates through convolutional learning and transformer-based contextual analysis. There were still cases of misclassification between tumor categories due to similar MRI intensities and image construction, but the hybrid model consistently provided diagnostic value and produced acceptable classification results across the MRI dataset.

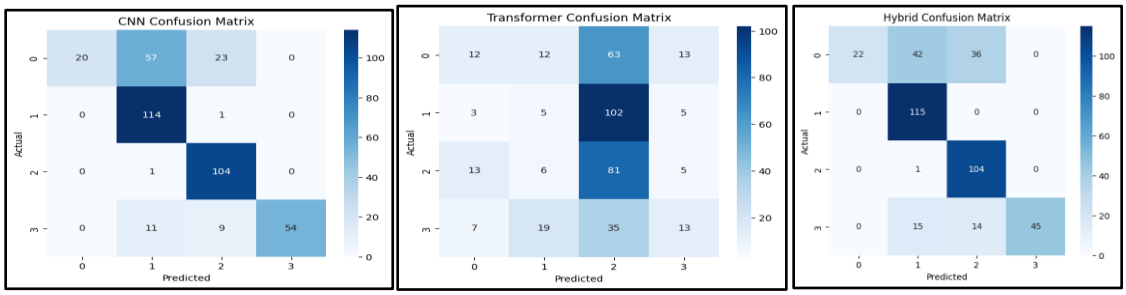


FIGURE 5. Confusion matrices of CNN, transformer, and hybrid CNN–transformer frameworks.

The discriminative performance was also tested at different classification thresholds using ROC-AUC analysis. The CNN-only model achieved an AUC of 0.89, the highest score, while the Hybrid CNN-Transformer model achieved an AUC of 0.84. The Transformer-only model achieved an AUC of 0.60, indicating reduced diagnostic performance.

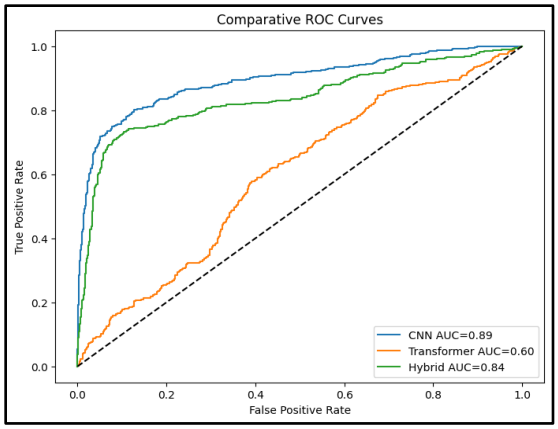


FIGURE 6. Comparative ROC-AUC curves of CNN, transformer, and hybrid models.

The overall assessment outcomes showed that CNN-based architectures still hold promise for MRI classification, and the proposed hybrid structure achieved competitive performance by combining the benefits of convenient feature extraction with the contextual learning provided by the Transformer.

2. COMPARATIVE ANALYSIS

The performance of the proposed Hybrid CNN-Transformer model was compared with CNN-only and Transformer-only models. This comparison aimed to determine whether the combination of convolutional feature extraction and Transformer-based contextual learning could yield better results in medical image diagnosis tasks. The CNN-only model achieved the highest classification accuracy and ROC-AUC among the tested frameworks. The result shows that a CNN has a strong capability to extract local spatial information from an MRI image. CNN architectures effectively capture tumor boundaries, texture patterns,and structural abnormalities that play a key role in tumor classification. Additionally, DenseNet121 was used to boost feature propagation and training stability. The performance of the Transformer-only model was significantly lower than that of the CNN and hybrid models. This result aligns with current medical imaging studies, in which using larger datasets can lead to more effective learning with Vision Transformers. Overall, the dataset used for the MRI was rather small, which limited the Transformer model's ability to generalize efficiently. In addition, it had limited ability to capture the tumor's local structures due to the absence of convolutional inductive biases.

The Hybrid CNN-Transformer is proposed to fuse the strengths of both architectures. CNN-based low-level and mid-level spatial features with Transformer encoder-based global contextual relationships learned with self-attention mechanisms. Though the hybrid model was slightly worse than the CNN baseline, its performance was competitive in classification, and it was also designed to be interpretable in terms of the learned contextual attention. These results were confirmed using a comparative ROC analysis. The Convolutional Neural Network (CNN) architecture achieves an ROC-AUC of 0.89, indicating strong discriminative performance. The Hybrid CNN-Transformer architecture demonstrated a consistently high AUC of 0.84, indicating that it achieved nearly the same diagnostic performance as the CNN architecture. In contrast, the Transformer architecture achieved an AUC of 0.60, indicating lower generalization due to the limited amount of training data from the MRI scans. In general, the comparative analysis showed that hybrid architectures can effectively combine the advantages of global Transformer contextual learning and local CNN feature extraction. The proposed framework thus provides a balanced solution to achieve both classification performance and interpretability in AI-based medical image diagnostic systems.

3. EXPLAINABILITY ANALYSIS

The proposed framework also integrates Explainable Artificial Intelligence (XAI) techniques, further enhancing transparency and interpretability in automated medical diagnosis. Grad-CAM visualization was

used to identify the regions of the image that contribute most to the model's predictions. In healthcare, explainability is crucial, as clinicians need interpretable evidence before implementing AI-powered diagnostic tools in clinical settings. The reason is that the CNN-only and Hybrid CNN-Transformer models were able to highlight clinically relevant tumor regions in MRI images in their Grad-CAM outputs. The heat maps showed areas of activation in red and yellow, with high activation in regions corresponding to the tumor, indicating that the models learned meaningful pathological features related to brain tumors.

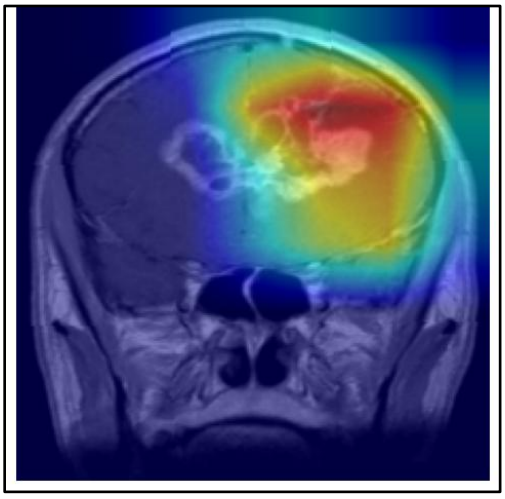


FIGURE 7. CNN Grad-CAM visualization.

CNN Grad-CAM visualization revealed highly localized activation patterns around the tumor masses. This suggests that the CNN architecture was able to capture some local spatial characteristics, such as tumor texture, boundaries, and structural abnormalities. The better quantitative performance of the CNN-only approach could be attributed to the strong localized attention.

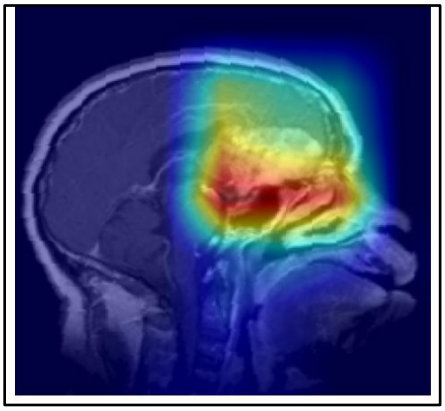


FIGURE 8. Hybrid CNN-transformer Grad-CAM visualization.

The Hybrid CNN-Transformer Grad-CAM output showed relatively wider distributions of activations around the anatomically adjacent areas. This indicates that the Transformer encoder captured more contextual information than just the local structures of the tumor. The hybrid framework thus integrated the local pathological feature extraction with the global contextual analysis using a self-attention mechanism. The explainability analysis confirmed that the proposed framework focuses on clinically relevant MRI regions rather than irrelevant background structures. This will enhance the trustworthiness and interpretability of the automated diagnosis process. Moreover, Grad-CAM visualization can support the use of hybrid AI models in clinical decision-support systems by providing visual explanations for diagnostic predictions.

4. DISCUSSION

The results of this study showed that deep learning architectures have outstanding diagnostic capability for brain tumor classification based on MRI images. Among the analyzed models, the CNN-only model achieved the highest classification accuracy and ROC-AUC. This suggests that CNNs are applicable to medical image analysis problems, even for moderately sized MRI datasets (Montaha et al., 2022). CNNs are particularly effective at learning localized spatial features, such as tumor boundaries, texture changes, and structural abnormalities, that play a crucial role in accurate tumor diagnosis (Jiang et al., 2023). The proposed Hybrid CNN–Transformer framework did not outperform the CNN-only model in quantitative results, but it did offer several salient points. The hybrid architecture achieved a balance between convolutional feature extraction and Transformer-based contextual learning, making it a success. The results from the Grad-CAM visualizations indicated that the hybrid model had a more broadly distributed activation across anatomical regions surrounding the tumor, reflecting better contextual understanding than the CNN-only approach. This augmented contextual learning can help to improve the clinical interpretability and diagnostic reasoning. The results for the Transformer-only model were relatively poor across all evaluation metrics. This is in line with a prior study that indicated that Vision Transformer models require large amounts of data to achieve optimal performance (Papa et al., 2024). The data used in this study were relatively limited, and the limited MRI data did not allow the Transformer model to effectively learn generalized representations. The explainability analysis demonstrated the importance of applying XAI techniques in medical imaging systems from a practical perspective (Fontes et al., 2024). The Grad-CAM visualizations were also found to agree with the evaluated models, as they directed attention to clinically relevant regions of the tumor rather than to unrelated artifacts within the images. Transparency is crucial in healthcare settings as it enables radiologists to make informed decisions before engaging AI-powered diagnostic tools (Oladele, 2024). This hybrid framework thus not only improves classification accuracy but also enhances clinical trustworthiness and interpretability. Limitations were found during this study that need to be addressed. The study involved only a relatively small amount of MRI data, which may limit the Transformer model's generalizability. The article has not taken into account other medical imaging modalities, such as CT or ultrasound, for tumors and has only considered tumor classification by MRI. Also, Transformer-based architectures increase computational complexity, thereby increasing training costs, compared to CNN models. In the future, there could be larger, multimodal medical imaging datasets and more elaborate Transformer models that could improve diagnostic performance. Explainability can also be further enhanced by other explainability methods such as SHAP or other attention visualization techniques. Additionally, transfer learning and multimodal fusion approaches could be explored to enhance generalization performance across medical imaging domains. Overall, the research showed that hybrid CNN–Transformer models are effective and explainable for automated medical image diagnosis. The proposed framework effectively merged local spatial feature extraction with global contextual learning, while maintaining explainability through Grad-CAM visualization. The results suggest that the field of clinical decision-making in radiology and medical imaging holds great promise for future AI-based clinical decision-support applications.

VI. CONCLUSION AND RECOMMENDATION

This study proposes and validates a Hybrid CNN–Transformer framework to find explainable classifiers for brain tumor classification based on MRI images. The architecture leveraged the local feature-extraction power of convolutional neural networks and the contextual learning power of the transformer architecture. Furthermore, to enhance the transparency and interpretability of diagnostic predictions, Explainable Artificial Intelligence techniques were applied using Grad-CAM visualization. Experimental results show that Deep Learning methods can provide useful assistance in medical image diagnosis. The CNN-only model was the best model in terms of classification performance (accuracy = 80.46%, ROC-AUC = 0.89). The proposed Hybrid CNN–Transformer framework achieved competitive results, with 74.37% accuracy and an ROC-AUC of 0.84, and offered an improved level of interpretability by enabling contextual attention learning. The Transformer-only model, on the other hand, achieved lower performance, suggesting that self-attention alone did not perform well for classification on the available MRI dataset. The Grad-CAM visualizations confirmed that both the CNN-only and Hybrid CNN–Transformer frameworks focused on

clinically relevant tumor regions during evaluation. CNN showed extremely localized activation around the tumor regions, whereas the hybrid model showed global activation across the context surrounding anatomical components. The results reveal that Transformer-based contextual learning improved interpretability and a more thorough understanding of MRI image characteristics.

The study successfully achieved all research objectives. The study included pre-processing and preparation of MRI images for analysis, implementation of CNN and Transformer in a single hybrid system, application of explainable AI (XAI) for model transparency, and use of widely adopted medical imaging performance measures to evaluate the framework. The research questions were answered using quantitative evaluation, comparison, and explanation assessment. In addition to the positive results, some limitations were identified. The MRI dataset was relatively small, which could limit the learning potential of Transformer-based models. Additionally, only MRI images were considered, and other imaging modalities, such as CT and ultrasound, were not used in the final-stage implementation. To improve effectiveness in medical diagnosis, future work can involve expanding the dataset with more images and greater diversity, adopting advanced Transformer models, and starting with multimodal medical imaging models. Based on the results obtained, it is suggested that future studies use larger datasets, improved multimodal fusion methods, and additional explanations such as SHAP and attention visualizations. The ability to integrate with clinical decision-support systems might also be considered, as this will aid in the deployment of the Healthcare system. Overall, the presented Hybrid CNN–Transformer framework proved effective for brain tumor classification and increased the interpretability of the model through explainable AI techniques. The framework offers a compelling approach to creating accurate, transparent, and clinically reliable AI-assisted medical diagnosis systems.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

1. Mohamed, E., Gaber, T., Karam, O., & Rashed, E. A. (2022). A Novel CNN pooling layer for breast cancer segmentation and classification from thermograms. Plos one, 17(10), e0276523. https://doi.org/10.1371/journal.pone.0276523
2. Alruwaili, A. J. A., Alhawiti, S. A. A., Alhawiti, R. A. H., ALshammari, S. S. R., Aldhuwalia, Y. I., ALOTAIBI, H. A. N., & saleh Alzaid, H. (2024). Advances in Medical Imaging for Early Diagnosis and Disease Management. *Saudi Journal of Medicine and Public Health*, *1*(2), 1584-1595. https://doi.org/10.64483/202412445
3. Al-Shoukry, S., & Musa, Z. B. (2024). An overview on effectiveness of activation functions in processing medical images. Al-Furat Journal of Innovations in Electronics and Computer Engineering, 3(2), 163-180. https://doi.org/10.46649/fjiece.v3.2.13a.22.5.2024
4. Applications of Transformers in medical image analysis: https://www.researchgate.net/figure/Applications-of-Transformers-in-medical-image-analysis-as-reviewed-in-this-work_fig5_358846126
5. Erben, A., Mayer, R., & Jacobsen, H. A. (2024). How can we train deep learning models across clouds and continents? An experimental study. *Proceedings of the VLDB Endowment, 17*(6), 1214–1226.
6. Fontes, M., De Almeida, J. D. S., & Cunha, A. (2024). Application of example-based explainable artificial intelligence (XAI) for analysis and interpretation of medical imaging: a systematic review. IEEE Access, 12, 26419-26427. Digital Object Identifier 10.1109/ACCESS.2024.3367606
7. GeeksforGeeks (2017). Introduction to Convolution Neural Network. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/introduction-convolution-neural-network/.

8. Nazir, M. I., Akter, A., Wadud, M. A. H., & Uddin, M. A. (2024). Utilizing customized CNN for brain tumor prediction with explainable AI. *Heliyon*, *10*(20), e38997.
9. Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Medical Imaging, 24(1), 1-19. https://doi.org/10.1186/s12880-024-01292-7
10. Guo, S., Wang, L., Chen, Q., Wang, L., Zhang, J., & Zhu, Y. (2022). Multimodal MRI image decision fusion-based network for glioma classification. Frontiers in oncology, 12, 819673. https://doi.org/10.3389/fonc.2022.819673
11. Henry, E. U., Emebob, O., & Omonhinmin, C. A. (2022). Vision transformers in medical imaging: A review. arXiv preprint arXiv:2211.10043. https://arxiv.org/pdf/2211.10043
12. Hou, Y., Wu, Z., Cai, X., & Zhu, T. (2024). The application of an improved DenseNet algorithm in accurate image recognition. Scientific Reports, 14(1), 8645. https://doi.org/10.1038/s41598-024-58421-z
13. Jiang, X., Hu, Z., Wang, S., & Zhang, Y. (2023). Deep learning for medical image-based cancer diagnosis. Cancers, 15(14), 3608. https://doi.org/10.3390/cancers15143608
14. *brain-tumor-mri-dataset* (2024). *Zenodo*, Version v2. https://doi.org/10.5281/zenodo.12735702 — this 2024 dataset record explicitly identifies the source as the Brain Tumor MRI Dataset from Kaggle.
15. Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: revolutionising accuracy and efficiency. Computer Methods and programs in biomedicine update, 5, 100146. https://doi.org/10.1016/j.cmpbup.2024.100146
16. Kumar, A. (2022). Deep learning for multi-modal medical imaging fusion: Enhancing diagnostic accuracy in complex disease detection. *Int J Eng Technol Res Manag*, *6*(11), 183. https://www.researchgate.net/profile/Anil-Kumar-541/publication/389988636_DEEP_LEARNING_FOR_MULTI-MODAL_MEDICAL_IMAGING_FUSION_ENHANCING_DIAGNOSTIC_ACCURACY_IN_COMPLEX_DISEASE_DETECTION/links/67db947ee62c604a0df5f6f4/DEEP-LEARNING-FOR-MULTI-MODAL-MEDICAL-IMAGING-FUSION-ENHANCING-DIAGNOSTIC-ACCURACY-IN-COMPLEX-DISEASE-DETECTION.pdf
17. Liao, L., Li, H., Shang, W., & Ma, L. (2022). An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. ACM Transactions on Software Engineering and Methodology (TOSEM), 31(3), 1-40. https://doi.org/10.1145/3506695
18. Montaha, S., Azam, S., Rafid, A. R. H., Hasan, M. Z., Karim, A., Hasib, K. M., ... & Mannan, Z. I. (2022). MNet-10: A robust shallow convolutional neural network model performing an ablation study on medical images to assess the effectiveness of applying an optimal data augmentation technique. Frontiers in Medicine, 9, 924979. https://doi.org/10.3389/fmed.2022.924979
19. Niu, J., Liu, S., Li, H., Zhang, T., & Wang, L. (2023). Grasp Detection Combining Self-Attention with CNN in Complex Scenes. *Applied Sciences*, *13*(17), 9655. https://doi.org/10.3390/app13179655
20. Nouis, S. C., Uren, V., & Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK. *BMC Medical Ethics*, *26*(1), 89. https://doi.org/10.1186/s12910-025-01243-z
21. Obuchowicz, R., Strzelecki, M., & Piórkowski, A. (2024). Clinical applications of artificial intelligence in medical imaging and image processing—A review. *Cancers*, *16*(10), 1870. https://doi.org/10.3390/cancers16101870
22. Oladele, O. K. (2024). AI-powered medical imaging: a comprehensive review of applications, benefits, and challenges. Google Scholar. https://www.researchgate.net/profile/Oluwaseyi-Oladele-3/publication/385074391_AI-Powered_Medical_Imaging_A_Comprehensive_Review_of_Applications_Benefits_and_Challenges/links/6713ddd8069cb92a811d0105/AI-Powered-Medical-Imaging-A-Comprehensive-Review-of-Applications-Benefits-and-Challenges.pdf
23. Papa, L., Russo, P., Amerini, I., & Zhou, L. (2024). A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking. IEEE transactions on pattern analysis and machine intelligence, 46(12), 7682-7700. https://ieeexplore.ieee.org/iel7/34/4359286/10508091.pdf
24. Qian, J., Li, H., Wang, J., & He, L. (2023). Recent advances in explainable artificial intelligence for magnetic resonance imaging. Diagnostics, 13(9), 1571. https://doi.org/10.3390/diagnostics13091571
25. Poudyal, A., Devkota, N., & Budha, G. (2024). MRI-based brain tumor classification using transfer learning: A comparative analysis. *International Journal of Computer Applications, 186*(40), 7–11.
26. Krishnan, P. T., Krishnadoss, P., Khandelwal, M., Gupta, D., Nihaal, A., & Kumar, T. S. (2024). Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer. *Frontiers in Neuroinformatics*, *18*, 1414925.
27. Shi, D., Liu, R., Tao, L., & Yuan, C. (2022, May). Heuristic dropout: An efficient regularization method for medical image segmentation models. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1101-1105). IEEE. DOI: 10.1109/ICASSP43922.2022.9747409
28. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11), 4793-4813. https://ieeexplore.ieee.org/iel7/5962385/9591206/09233366.pdf

29. Toh, Y. (2024). Ultrasound versus magnetic resonance imaging as first-line imaging strategies for rotator cuff pathologies: a comprehensive analysis of clinical practices, economic efficiency, and future perspectives. *Cureus*, *16*(4). https://doi.org/10.7759/cureus.59231
30. Xia, K., & Wang, J. (2023). Recent advances of transformers in medical image analysis: a comprehensive review. MedComm–Future Medicine, 2(1), e38. DOI: 10.1002/mef2.38
31. Anghel, C., Grasu, M. C., Anghel, D. A., Rusu-Munteanu, G. I., Dumitru, R. L., & Lupescu, I. G. (2024). Pancreatic adenocarcinoma: Imaging modalities and the role of artificial intelligence in analyzing CT and MRI images. *Diagnostics*, *14*(4), 438. https://doi.org/10.3390/diagnostics14040438
32. Zhang, C., Deng, X., & Ling, S. H. (2024). Next-gen medical imaging: U-Net evolution and the rise of transformers. *Sensors*, *24*(14), 4668. https://doi.org/10.3390/s24144668
33. Zhao, S. S., Feng, X. L., Hu, Y. C., Han, Y., Tian, Q., Sun, Y. Z., ... & Wang, W. (2020). Better efficacy in differentiating WHO grade II from III oligodendrogliomas with machine learning than radiologists' reading from conventional T1 contrast-enhanced and fluid-attenuated inversion recovery images. BMC neurology, 20(1), 48. https://doi.org/10.1186/s12883-020-1613-y
34. Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., ... & Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 109(5), 820-838. doi:10.1109/JPROC.2021.3054390