ISSN: 2093-5374 / E-ISSN: 2093-5382

Vision Transformer Based Cross Disorder MRI Screening with Interpretable Saliency Guided Decisions

Ratnakala Patil[1], and Dr. Sachinkumar Veerashetty[2*]

[1]Assistant Professor, Department of Computer Science & Engineering, Sharnbasva University, Kalaburagi, Karnataka, India. ratnakala@sharnbasvauniversity.edu.in, https://orcid.org/0009-0006-9553-4487

[2*]Professor, Department of Computer Science & Engineering, Sharnbasva University, Kalaburagi, Karnataka, India. sveerashetty@sharnbasvauniversity.edu.in; sveerashetty@gmail.com, https://orcid.org/0000-0001-8217-1388

Received: February 28, 2026; Revised: April 06, 2026; Accepted: May 22, 2026; Published: June 30, 2026

Abstract

The diagnosis at an early stage for the neurological problems like Alzheimer's disease, epilepsy, and brain tumors plays a critical role in increasing the chances of recovery from these diseases. However, traditional CNN techniques used for identifying such neurological problems in MRI images suffer from the limitations of understanding the long-range dependencies and context relations among various kinds of brain diseases. This paper proposes ViT-CrossMRI, which uses the vision transformer to perform cross-disorder MRI analysis and generates saliency guided by interpretations. It makes decision-making processes more clinically meaningful and transparent. The model was tested experimentally using multi-center heterogeneous datasets, where superior results were obtained in terms of several different measures, such as accuracy (93.12%), precision (92.76%), recall (92.30%), F1-score (92.53%) and AUC-ROC (95%). Balanced sensitivities and specificities of 92.30% and 93.45%, respectively, indicate that both positive and negative cases are identified effectively. Compared to baseline models such as CNN, ResNet-50, and LSTM, this approach demonstrates superiority in relation to the general ability to use global context and attention in cross-disorder diagnosis. From the computational point of view, reasonable inference time (145 ms for each MRI scan) and low parameter number (86.2M) and memory usage (6.8 GB) can be noted.

Keywords: Vision Transformer (ViT), Cross-Disease Screening, Brain MRI, Alzheimer's Disease, Epilepsy, Brain Tumor Detection, Explainable AI (XAI).

1 Introduction

Alzheimer's disease, epilepsy, and brain tumors are some examples of highly prevalent brain disorders that are progressive and hard to diagnose; these are among the most challenging aspects of worldwide health conditions (Do & Hill, 2023). Accurate diagnosis is crucial for better prognosis, development of treatment approaches, and lower costs of medical care. Magnetic Resonance Imaging (MRI) is an effective tool for identification of brain anomalies that does not involve any invasion, is noninvasive, and is informative in respect to both the structure and pathology of the brain. However, reading of MRI

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 593-613. DOI: *10.58346/JOWUA.2026.I2.033*

*Corresponding author: Professor, Department of Computer Science & Engineering, Sharnbasva University, Kalaburagi, Karnataka, India.

images requires much time and clinical expertise; moreover, it varies from person to person (Daraban et al., 2024). These difficulties make the development of efficient and reliable computer-assisted methods for diagnosing diverse brain diseases indispensable (Wortmann, 2012).

Deep learning techniques, especially the Convolutional Neural Network (CNN), have proven to be extremely effective in processing medical imaging data in recent times (Mora-Rubio et al., 2023; Iqbal et al., 2023). The use of convolutional filters and hierarchical representation in the CNN model has been effective in extracting local spatial features (Bravo-Ortíz et al., 2021; Wang et al., 2025). Despite the success achieved, CNNs are mainly dependent on localized receptive fields, making them less efficient in analyzing global contextual dependencies and distant correlations in brain MRI data (Alahmed & Al-Suhail, 2025; Xin et al., 2023). There are various neurological diseases that show slight structural changes in the brain at different parts of the brain. Therefore, modeling distant interactions between anatomical structures is critical in diagnosing remote diseases. Vision Transformer (ViT) is quite a recent development and poses significant competition to the CNNs in computer vision applications (Zhao et al., 2023; Khan et al., 2023). ViTs utilize the self-attention mechanism that splits the image into patches while understanding global interactions between the patches. Such self-attention strategy enables the model to weigh the significance of each patch with respect to other parts, leading to better capture of long-range correlation than the regular convolution mechanism (Jahangir et al., 2024; Khojaste-Sarakhsi et al., 2022).

The other significant issue in implementing artificial intelligence in healthcare is interpretability (Tanveer et al., 2020). The process of clinical decision-making requires openness, responsibility and trust. Black-box models that produce uninterpretable predictions are usually not applicable to the medical domain. To address this problem, explainable AI (XAI) methods have been proposed to understand and visualize the model's decisions. Saliency maps in medical imaging can also be used to identify areas of high predictive value and assists clinicians in determining whether the model emphasizes clinically significant anatomical features. The proposed research recommends a Vision Transformer-based system for cross-disease screening of brain disorders using MRI images (Rifat et al., 2025). The framework incorporates a layer of saliency-directed interpretability that produces attention-based explanations for every prediction. Compared with single-disease classification models, the proposed model is designed to be generalized across various neurological diseases, such as Alzheimer's disease, epilepsy, and brain tumors (Scheltens et al., 2021) (Association, 2015). The system is expected to enhance diagnostic precision and clinical interpretability by incorporating global feature modeling, with clearer visualization of decisions (Sharma & Mandal, 2022).

Key Contributions of the Research

- The development of a ViT-based unified architecture for multi-disease brain MRI classification.
- The integration of a saliency-guided explanation mechanism to enhance transparency and clinician trust,
- Comprehensive evaluation demonstrating cross-disease generalization capability.
- The proposed approach has the potential to support radiologists and neurologists in early-stage diagnosis, reduce diagnostic variability, and advance the practical deployment of explainable AI in neuroimaging applications.

The organization of this paper will continue as follows. Section 2 will cover the literature review on the state-of-the-art deep learning-based approaches for brain MRI analysis, deep learning-based disease classification models using CNN, Vision Transformers development progress, and current explainable

AI solutions for medical images. Section 3 describes the methodology for the project, which includes vision transformer structure, data pre-processing, model training for various diseases, and the proposed saliency-based explanation method. Section 4 provides an overview of the results obtained from the experiments performed, describing the datasets used, implementation specifics, performance metrics applied, and the comparison of the proposed approach with baseline models. Finally, the discussion section in Section 5 will provide a brief overview of the experimental findings along with the key strengths of the model, generalization capabilities for various brain disorders, clinical relevance, and limitations of the approach.

2 Literature Review

ViTs and Convolutional Vision Transformers (CViTs) for Alzheimer's disease detection based on neuroimaging data. For their review of literature, which is used 68 scientific articles published in reliable scientific databases and classified according to architecture, multimodal fusion approaches, data sets, and methodologies. According to the authors, one of the trends in this research field is the increased prevalence of hybrid architectures combining the use of transformers with convolutional layers to provide a balance between local feature extraction and modeling dependencies within the image as a whole. In addition, several areas for further investigation have been identified, such as lack of consistent validation practices, limited use of external datasets, and underrepresentation of longitudinal studies. Another important aspect mentioned in the paper is the use of XAI approaches and lightweight transformer-based architectures, which would allow for the implementation of AI (Bhandarkar et al., 2024).

In this describes the Vision Transformer used in their framework identifies the higher-level spatial features of each of the 2D slices while the time series transformer captures the dependencies between the slices with the assumption that the MRI volume is a sequence. In addition, which is adopted dual transfer learning technique to minimize the number of required data. Which is where able to achieve higher classification accuracy compared to conventional CNN-based algorithms, making it evident that transformers have a high level of efficiency (Khatri & Kwon, 2024).

The different interpretability techniques that enhance the transparency of clinical decision-making, including Grad-CAM and attention visualization. The paper concluded that transformer models are promising for diagnostic accuracy, but that there are issues of data scarcity, computational constraints, and a lack of benchmarking standards (Alp et al., 2025).

The general applicability of Vision Transformers across a wide range of medical images, including classification, segmentation, reconstruction, and object detection. In their work, the authors focused on the benefits of transformers in capturing the most important features of the spatial dependence of objects over long distances via self-attention, thereby enabling better depiction of spread-out pathological patterns. The review divided transformer-based models based on the imaging modalities of MRI, CT, and ultrasound, and compared their performance in various organs and diseases. The authors have emphasized the issues of high computational cost, large data requirements, and low interpretability, and have proposed a hybrid CNN-transformer design as a feasible way to enhance robustness and efficiency in medical settings (Hosny & Mohammed, 2025).

A fine-tuned Vision Transformer model can effectively classify brain tumors from MRI scans. Which is used transfer learning in their study to extend a pre-trained ViT architecture to medical imaging data, enabling them to perform well with a small dataset. The model captured global spatial relationships in brain structures, achieving higher classification accuracy than traditional CNN methods. The study

provided a robust background on pure transformer-based tumor classification models and highlighted the potential of ViTs to perform well without complex hybrid adjustments. The authors further indicated that larger datasets from multiple institutions need to be validated to enable broader clinical application (Fang et al., 2023).

Table 1: Comparative analysis of vision transformer-based brain MRI studies

Ref No	Methodology	Dataset / Simulation Environment	Research Gap
(Bhandarkar et al., 2024)	Vision Transformer and Convolutional Vision Transformers (CViTs).	68 peer-reviewed studies from major scientific databases (neuroimaging-based Alzheimer's research)	Inconsistent validation protocols, limited external dataset testing, lack of longitudinal progression analysis, and insufficient lightweight XAI-enabled transformer models
(Khatri & Kwon, 2024)	Hybrid framework combining Vision Transformer (ViT) for spatial feature extraction and Time-Series Transformer	Structural MRI datasets for Alzheimer's disease classification	Focus limited to Alzheimer's disease; lacks cross-disease evaluation and clinical explainability integration
(Alp et al., 2025)	Survey of Vision Transformer-based brain tumor models integrated with Explainable AI (Grad-CAM, attention visualization)	MRI-based tumor diagnosis studies across supervised and unsupervised frameworks	Data scarcity, high computational cost, absence of standardized benchmarking, and limited real-time deployment feasibility
(Hosny & Mohammed, 2025)	Comprehensive review of Vision Transformers for classification, segmentation, detection, and reconstruction across medical imaging modalities	MRI, CT, and ultrasound datasets across multiple organs and diseases	High computational requirements, large-scale data dependency, and limited interpretability in clinical workflows
(Fang et al., 2023)	Fine-tuned pre-trained Vision Transformer	Brain tumor MRI datasets (single-modality clinical dataset)	Requires validation on multi-institutional datasets; lacks cross-pathology generalization and interpretability mechanisms
(Asiri et al., 2023)	Vision Transformer-based predictive modeling for early treatment response in brain metastases	MRI data of patients with brain metastases under treatment	Focused on treatment response prediction only; not evaluated for multi-disease screening or general diagnostic classification
(Volovăț et al., 2025)	Comparative study between Vision Transformers and CNN-based transfer learning models (VGG, ResNet) with Grad-CAM explainability	Brain disease MRI datasets for classification	Limited to performance comparison; lacks unified transformer-based cross-disease architecture
(Sarker et al., 2024)	ViT-B-16 architecture for MRI brain tumor classification with attention visualization	Brain tumor MRI datasets compared against CNN baselines	Focus restricted to tumor classification; lacks multi-disease screening framework and saliency-guided interpretability optimization

In this explored the application of vision transformers in predicting early treatment responses among brain metastasis patients using MRI images. Unlike the conventional approach of classification models, this study sought to develop prediction models for customized treatment planning. This was facilitated

by the attention model within the architecture, which allowed for the identification of imaging features associated with treatment response (Asiri et al., 2023). The findings indicated that the predictive model proved highly reliable, even under class imbalance conditions, which explains why vi-t's can be applied with clinical data. It was concluded that transformers-based predictive models would guide oncologists in making therapeutic decisions and customization of medicine. Vision Transformer models and transfer learning approaches including VGG and ResNet for identifying diseases from MRI images of the brain. It has been found that Vision Transformers perform better than CNN-based transfer learning models in terms of learning global dependency structures across different parts of the brain. Also included in the experiment are the techniques used to explain the model predictions, for example, using Grad-CAM. This helps in achieving more transparency and making the model clinically interpretable, which makes transformers highly suitable for use in medical imaging problems. The researchers clearly suggest that Vision Transformers are crucial for implementing AI in healthcare (Volovăț et al., 2025).

The ViT-B-16 architecture for the classification of brain tumors in MRI and compared it with traditional CNN models. According to their findings, their accuracy and feature representation capability were much higher with transformer-based attention mechanisms. The paper noted that ViT-B-16 is highly effective at handling high-dimensional MRI data by capturing long-range interactions among image patches. The authors also emphasized that attention visualization can enhance interpretability in clinical decision-making. The study has shown that transformer architectures have the potential to be a strong alternative to CNNs for neuroimaging tasks, enabling the development of more sophisticated cross-disease diagnostic systems (Sarker et al., 2024).

As table 1 demonstrates, Vision Transformers (ViTs) are becoming increasingly popular in brain MRI analysis, particularly for detecting brain tumors and Alzheimer's disease. Transformer- or hybrid-CNNViT models are popular among many researchers due to their ability to capture global image relationships, unlike traditional CNNs. Some of them used transfer learning to process small medical datasets and were explainable AI mechanisms, such as Grad-CAM or attention maps, to enhance clinical transparency. Nevertheless, most publications focus on a single disease and are not intended for screening multiple brain diseases. Besides that, there are still problems with low external validation, high computational cost, and absence of standardized testing. Such gaps indicate the need for a generalized, interpretable transformer-based framework for cross-disease brain MRI screening.

Research Gap

While CNNs and Vision Transformer approaches have proven successful in MRI-based disease diagnosis, these methods suffer from limitations such as poor cross-disease generalization, limited interpretability to assist clinicians in making better diagnoses, and dependence on datasets collected from a single institution. This paper aims to provide a new cross-disease MRI analysis approach that leverages globally learned features, saliency visualization, and strong cross-disease generalization across heterogeneous data.

3 Methodology

3.1 Overall Architecture of Proposed Methodology

In figure 1 showcases the full pipeline of a cross-disorder MRI scan screening system based on a Vision Transformer (ViT) architecture that allows for interpretable and saliency-driven decision making.

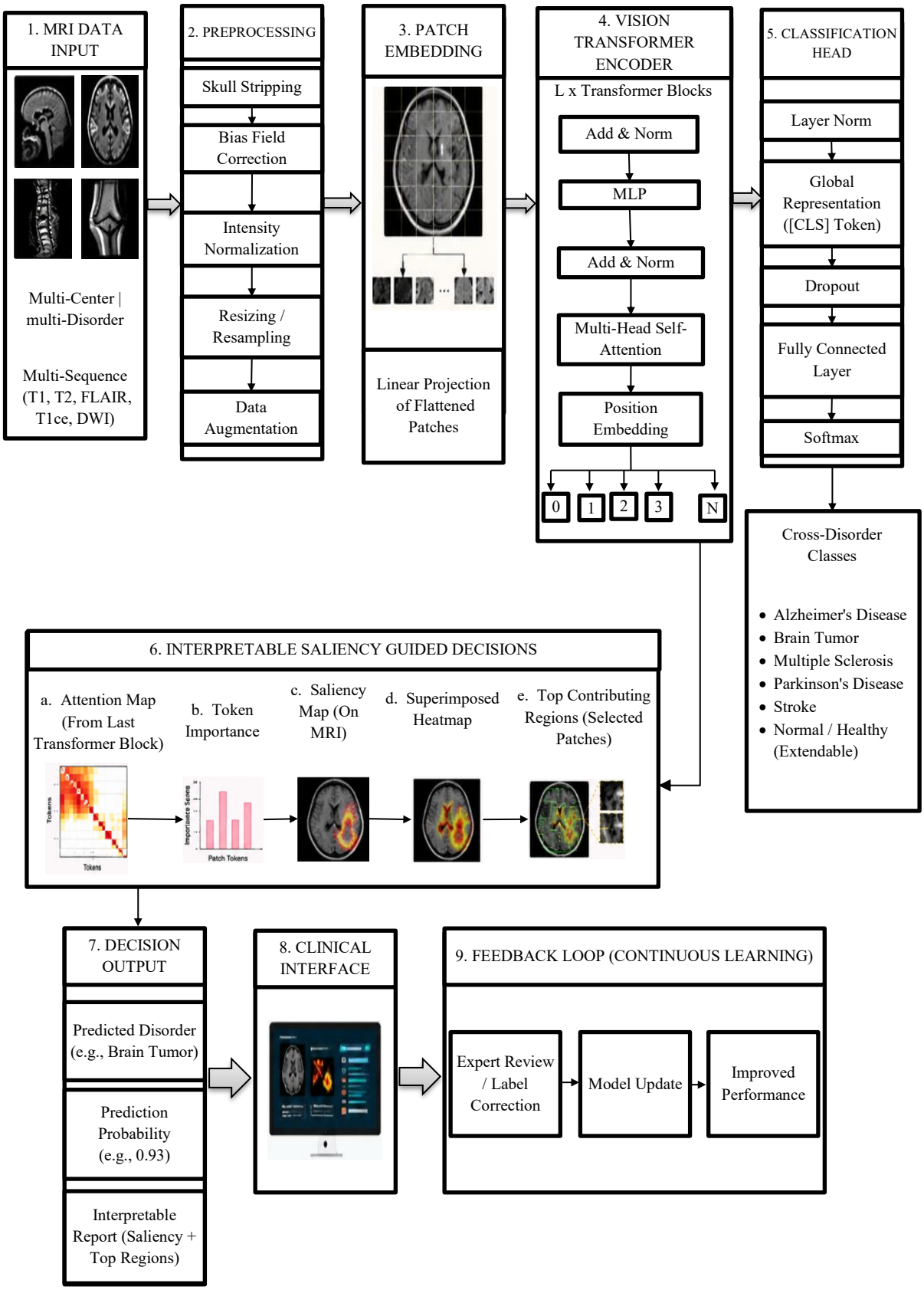

Figure 1: Overall architecture of the proposed methodology

Multicenter and multisource MRI scans undergo preprocessing techniques such as skull stripping, intensity normalization, bias correction, resizing, and data augmentation. Next, images are split into patches for linear projections before being input into the ViT Encoder. This encoder uses multi-head self-attention and a transformer-based architecture to capture global contextual features. Meanwhile, clinical features are independently encoded using an MLP before fusing together with imaging features via cross-modal attention. Predictions are generated by multiple types of classifiers, including GBM, GLM, ResNet, DenseNet, RF, and SVM classifiers. The resulting output is a prediction of the probabilities of multiple neurological disorders.

3.2 Working Principle of Vision Transformer for Saliency Guided Decisions

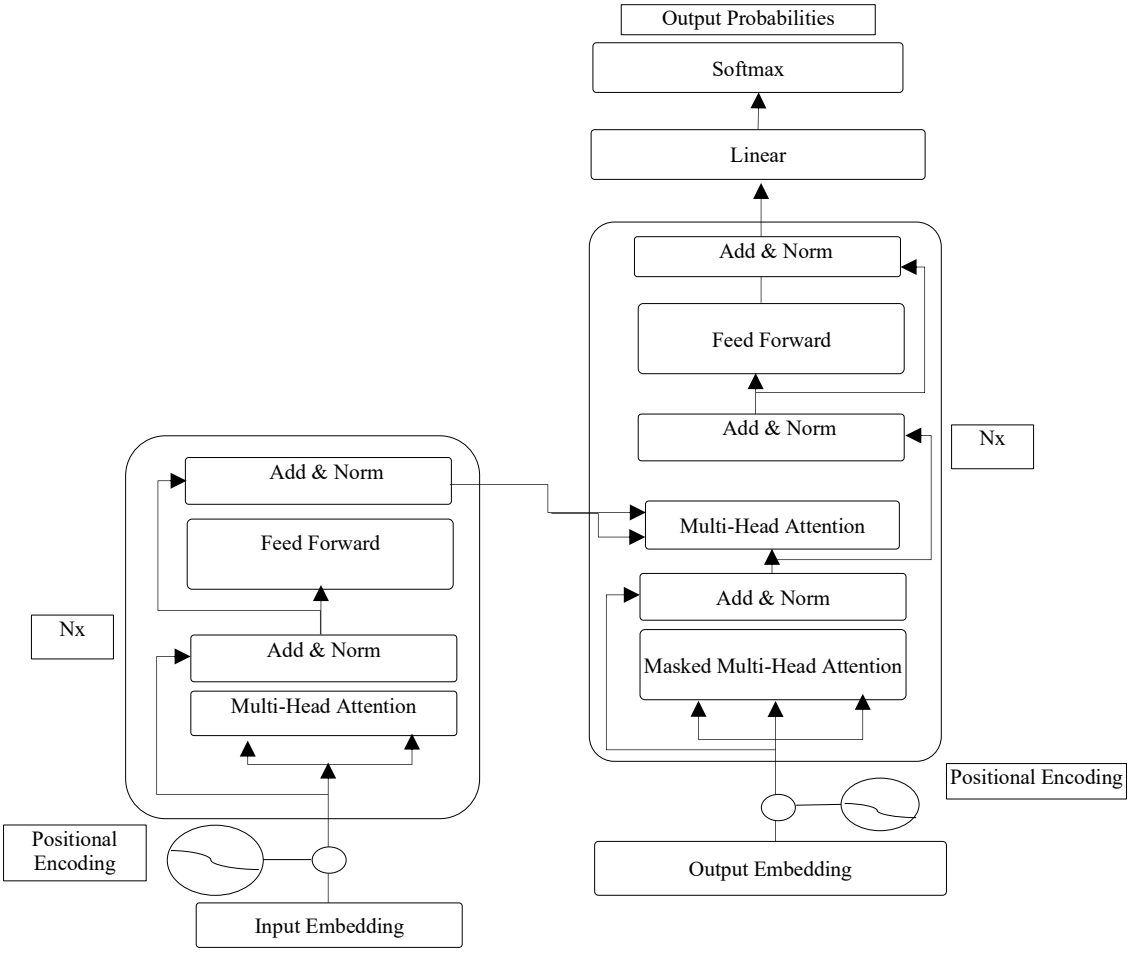

Figure 2: Working principle of vision transformer for saliency guided decisions

In figure 2 Shows the depiction of the architecture of a transformer model, particularly the encoder-decoder architecture that applies to sequence-to-sequence tasks such as language translation. In the depiction, on the left-hand side is the encoder part of the architecture where the input tokens are turned into embeddings and positional encodings are added to ensure that there is ordering maintained. Then, there are several stacked layers (N×), each with multi-head attention and a feed-forward network, all wrapped around an Add & Norm connection mechanism to stabilize learning. On the right side of the figure, the decoder process starts by converting the output embeddings, preventing any information leakage. Masked multi-head attention ensures that the model attends only to previous tokens while the

other multi-head attention is applied on the output from the encoder. The equation above shows that when you introduce a scale factor, the exploding gradient issue is solved. If the input is too large, the softmax activation function generates a very tiny gradient, which causes the learning process to become slow. In this case, d is the dimension of the attention head, as we shall see later on in the section.

The attention score matrix is used to represent the attention score. Here, each row of the matrix represents the attention of the particular embedding on all the other embeddings. This process is commonly called self-attention. The importance of the information is now weighted by the attention score and is combined to obtain the enriched embedding (Equation 1).

$$Enriched\ Embeddings(Q,K,V) = Softmax\left(\frac{Q^T K}{\sqrt{d}}\right)v \tag{1}$$

Attention Modules

As can be seen from the above, the analysis of attention weights is not an appropriate approach to explain ViT architecture. The most informative part of the architecture is the final layer, which at the same time is the most complex one. In order to understand the architecture better, Abnar and. Which is established a relationship between the last layer of the model and input tokens by introducing new disentangled attention weights A_{roll}. This method utilizes the attention weights of each layer and computes the information flow in the network. These methods are based on the same assumptions, using the same raw attention weights A^l from each layer, but differ in how which is calculate this flow characterized by the disentangled weights A_{roll}.

$$A_{res}^{l} = \frac{1}{2}A^l + \frac{1}{2}l \tag{2}$$

From the above equation (2) describes the $\frac{1}{2}$ used for balance the contribution of attention mechanism based on residual connections. The weight must be normalized through the transformations.

$$A_{roll}^{(L-t-1)} = A_{res}^{(L-t-1)} A_{roll}^{(L-1)} \tag{3}$$

From the above equation (3) describes the $A_{res}^{(L-t-1)} A_{roll}^{(L-1)}$ and $i \in \{0,1,\ldots.L-1\}$.

Layer-wise Relevance Propagation, on the other hand, is a method to compute the relative relevance of neurons at a certain point within a neural network in relation to a neuron at another point. It works by decomposing the model's decision into a sum of relevance scores originating from all previous layers, recursively up to the input layer. This is done based on the principle of conservation, whereby whatever was received by the neuron is entirely passed to the next layer downwards, irrespective of where the layer appears.

$$R_j = \sum_k \frac{Z_{jk}}{\sum_j Z_{jk}} R_k \tag{4}$$

According to equation (4), it can be understood that the partial LRP method uses the LRP method to find out how the contributions from different attention heads affect the model predictions rather than taking the average of attention heads into consideration. But the partial LRP method tries to determine the relevance of the attention heads to visualize them and prune irrelevant heads. However, it does not try to establish the relationship between the relevance of attention heads and the relevance of the tokens on them. So, it just acts as a medium for complete interpretation from the prediction results and the input tokens.

$$\nabla A^l = A^l \left(\frac{\partial S_c}{\partial A^l} \right) \quad (5)$$

From the above equation (5) describes the ∇A^l defined the account of many components with a transformer encoder to matrix v with normalization and residual connections.

$$R^c = R^c + \nabla A^l R^c \quad (6)$$

Equation (6) explains that a Markov chain is completely determined by its transition matrix. This matrix gives the probability of moving from one state to another for each state. In this case, at each block, the output embeddings are treated as the states of the Markov chain. The transition matrix is built using the attention weightsA^l.

3.3 Data Acquisition and Preprocessing

The research proposal will use a structural MRI dataset from a publicly available neuroimaging library and clinically validated resources for various brain diseases, including Alzheimer's disease, epilepsy, and brain tumors, to allow cross-disease screening within a single framework. To ensure generalization, heterogeneous datasets from various scanners, acquisition plans, and resolutions are presented. Stratified sampling is used to split the data into training, validation, and test sets to maintain class balance across all disease groups. First, non-brain tissue is removed via skull-stripping, and the data are normalized to mitigate inter-scanner differences and promote uniformity. The sizes of all MRI volumes are adjusted to match a specific resolution (e.g., 224 X 224), which is suitable for processing with the Vision Transformer. With respect to 3D MRIs, slicing is another stage that transforms three-dimensional images into two-dimensional while maintaining anatomical continuity. Some data augmentation methods include rotations, flips, scaling, and contrast adjustments. Another type of pre-processing includes z-score normalization, which promotes stability in the distribution of intensities. Finally, the class imbalance is mitigated by applying weighted sampling techniques. In addition, all images are cut up into non-overlapping patches of a certain size and are then flattened and linearly embedded before being processed by the transformer encoder.

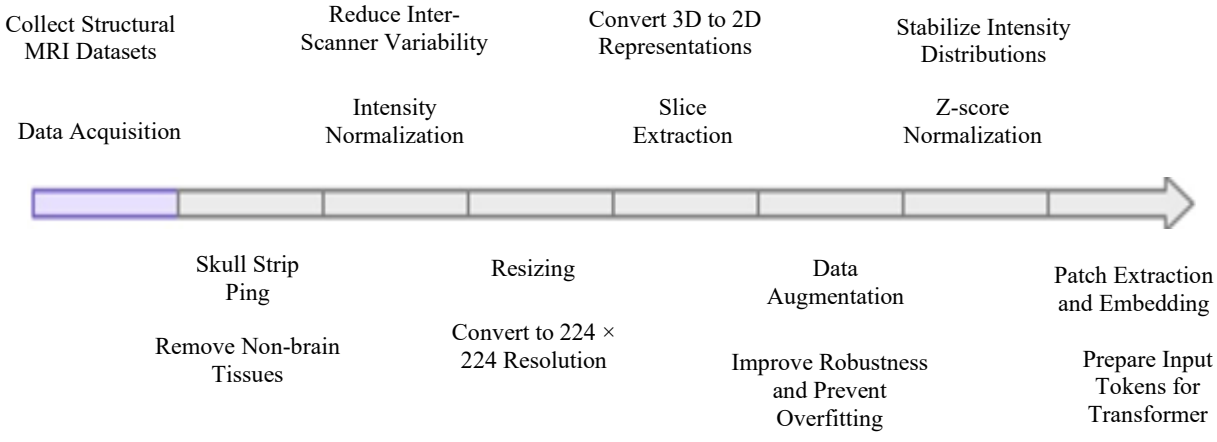

Figure 3: Structural MRI data preprocessing and transformer-based input preparation pipeline

The figure 3 shows the whole preprocessing workflow of structural MRI images prior to their input into a machine learning model powered by transformers. The steps involved include the beginning of the process with data acquisition in which structural MRI datasets are obtained. Normalization of intensity is made to minimize inter-scanner variability and enhance consistency and skull stripping is

done to eliminate non-brain tissue and keep only the relevant anatomical features. The volumetric 3D MRI scans are further reduced to 2D representations by slice extraction allowing effective processing. Every slice is re-sized to 224x224 resolution to ensure that it is compatible with a few of the common deep learning designs. Data augmentation techniques are used to make the data better and to avoid overfitting. This step is followed by Z-score normalization which normalizes the intensity distributions among the samples to enhance model convergence. Lastly, patch extraction and embedding are carried out to convert the processed images into input tokens that could be transformer models. Such a preprocessing workflow allows data consistency, ensures a better representation of the features, and facilitates the training of the models in a stable and reliable way.

3.4 Vision Transformer-Based Cross-Disease Classification Framework

The suggested classification framework is based on a Vision Transformer (ViT) model that is aimed at screening brain disorder in a single unified model. As compared to conventional CNNs which mainly extract local spatial-based features by using convolution operations, the Vision Transformer uses self-attention to incorporate global dependencies in the entire MRI image to better identify distributed pathological patterns. Each MRI image that has been preprocessed is split into small, fixed-size pieces. To keep track of the spatial arrangement of these pieces, positional embeddings are added. A special token called [CLS] is also included to capture overall context from all the pieces. The self-attention mechanism calculates the correlations among all patches at once, and this mechanism is what makes the model able to detect the hidden structural abnormalities of far brain parts. The last output of the classification token is input to a fully connected layer which is then followed by a softmax activation to classify probability in the form of Alzheimer disease, epilepsy, brain tumor, and normal conditions. Transfer learning is utilized by initializing the model with pre-trained weights and the fine-tuning is implemented on MRI datasets with the Adam optimization of categorical cross-entropy loss and dropout and early stops are used to optimize the generalization performance.

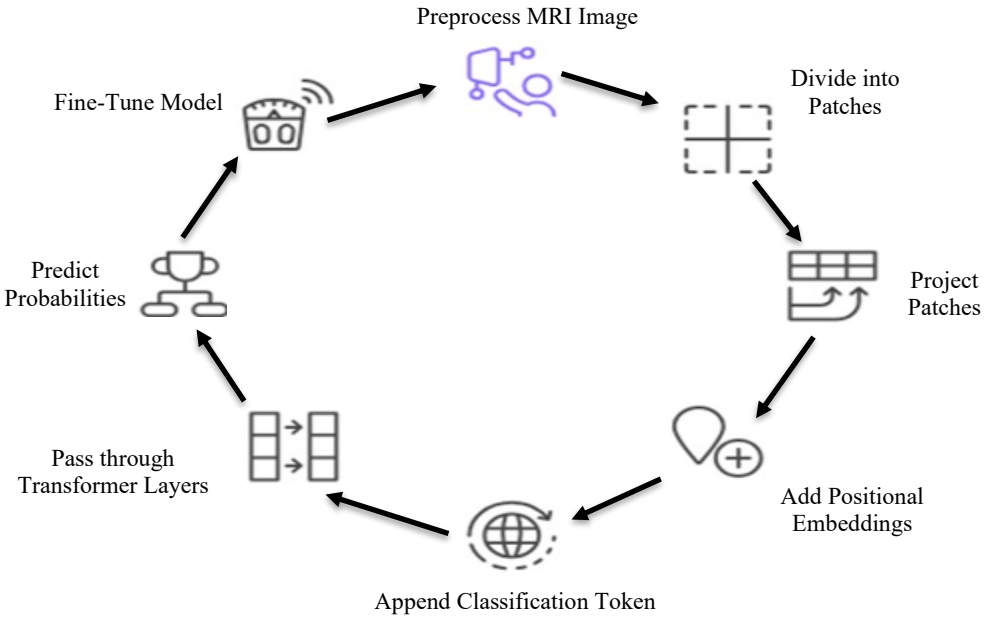

Figure 4: Vision transformer-based MRI classification workflow

The figure 4 shows the entire process of a Vision Transformer (ViT)-based architecture to classify MRI images. It takes a step of preprocessed MRI image into fixed-size patches to provide transformer-based processing. These patches are then mapped into embedded feature vectors whereby spatial image regions are tokenized. Position embeddings are embedded to retain the spatial information, which would be lost otherwise when patch flattening is applied. A classification token is added to the sequence, which is used as a universal depiction of final prediction. The token sequence is then forwarded through several layers of transformer encoders at which self-attention mechanisms are used to learn the global contextual relations among image-patches. The model gives prediction probability as a classification head. Lastly, the fine-tuning of the entire architecture is performed that aims to maximize performance in the particular task of the MRI classification. This built pipeline helps to effectively learn features on a global scale, neighbors better understand the context, and makes a more accurate prognosis of the diagnosis.

3.5 Saliency-Guided Interpretability Mechanism

A saliency-based interpretability module is incorporated into the suggested Vision Transformer framework to guarantee clinical transparency and reliability to give visual explanations to model predictions. The interpretability mechanism is based on the attention-based visualization and gradient-based saliency mapping methods to determine brain regions that have a significant impact on classification results. The attention weights produced in the multi-head self-attention layers are obtained and further summed over layers to produce the attention heatmaps that reveal the relationship between adjacent patches and the global focus areas. Besides that, gradient-based saliency analysis calculates the gradient of the predicted score of the classes with the input pixels, emphasizing areas in which subtle variations in intensity play a significant role in the prediction. These saliency maps are superimposed on the original MRI images in order to provide intuitive visual explanations that can help clinicians to comprehend the reasoning process of the model. The framework is designed in such a way that highlighted areas relate to significant pathological areas as tumor masses, cortical atrophy, or abnormal structural changes. To ensure reliability, quantitative assessment of the saliency localization performance is performed and the interpretability module is implemented to perform the task, without causing a significant rise in the complexity of computations. Combining this layer of saliency-driven explanation with the existing system, the proposed one improves the credibility, promotes the informed clinical choices, and allows the real application of the system in the field of the multi-disease brain MRI screeners.

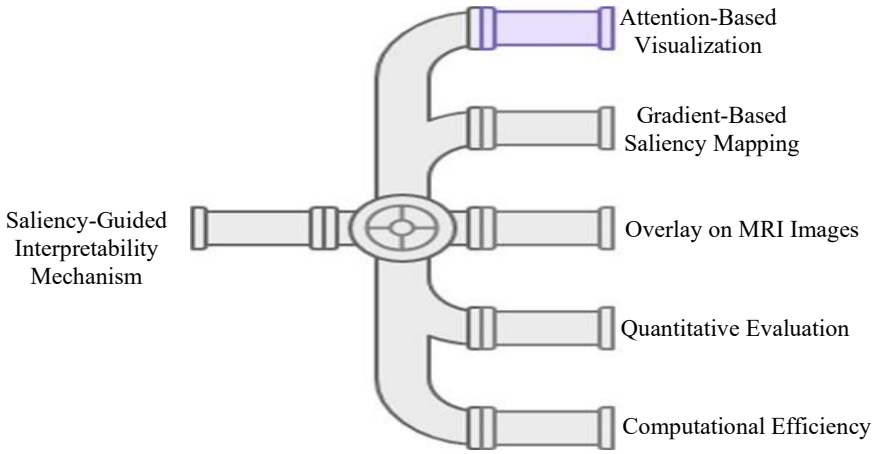

Figure 5: Saliency-guided interpretability framework for MRI classification

The figure 5 introduces a saliency-directed interpretability approach to deep learning models that is aimed to increase the transparency of the models based on MRI. The framework combines attention-based visualization and gradient-based saliency mapping methods to detect and draw attention to the most important areas that affect model predictions. These saliency maps are superimposed on MRI images to give understandable visual descriptions of the decision-making criterion in the model, which allows clinicians to check whether the model highlights diagnostically relevant areas. The framework also involves quantitative assessment in order to determine the reliability and consistency of interpretability outputs, such that the highlighted regions have any meaningful association with pathological aspects. Also, computational efficiency is felt to ensure realistic or large-scale clinical applicability. All in all, this interpretability mechanism enhances the transparency of the model, the establishment of clinical trust, and the creation of a more reliable and interpretable MRI classification system.

3.6 Multi-Head Self-Attention Mechanism

$$Attention(Q,K,V) = Softmax((QK^T)/Sqrt(d_k))V \quad (7)$$

The equation (7) shows self-attention mechanism computes the relationship between image patches by comparing Query (Q), Key (K), and Value (V) matrices derived from patch embeddings. The dot product between Q and K is scaled by the square root of the key dimension dk to stabilize gradients. The softmax is used to convert the attention scores into probabilities that are subsequently multiplied by V to produce context-aware feature representations. The equation is important as it allows the model-to-model
long-range spatial dependencies throughout the whole MRI image that is vital in identifying distributed brain abnormalities.

Cross-Entropy Loss Function for Multi-Class Classification

$$L = -sum(y_i log(pi)) \quad (8)$$

In equation (8), Categorical cross-entropy loss compares the true class labels yi and the classification probabilities pi generated by the softmax layer. The loss is more severe when the probability of the prediction in the correct class is low. Reducing such loss with training is beneficial in making the Vision Transformer acquire discriminatory features to correctly classify the presence of various brain disorders and abnormalities like Alzheimer's disease, epilepsy, brain tumors, and even a normal brain.

3.7 Proposed Algorithm

Algorithm: Vision Transformer-Based Cross-Disease MRI Screening

Input: MRI Dataset D

Output: Predicted Brain Disorder Class and Saliency Map

1: Load MRI dataset D

2: Perform preprocessing:

a. Skull stripping

b. Intensity normalization

c. Resize images to fixed size

d. Apply data augmentation

3: Split dataset into Training, Validation, and Test sets

4: For each MRI image I in Training set:

a. Divide I into fixed-size patches

b. Flatten and embed patches

c. Add positional encoding

d. Pass embeddings through Transformer encoder

e. Extract CLS token representation

f. Compute class probabilities using Softmax

g. Calculate Cross-Entropy loss

h. Update model weights using Adam optimizer

5: Validate model performance on Validation set

6: Test trained model on Test set

7: Generate attention maps and gradient-based saliency maps

8: Overlay saliency map on original MRI image

9: Return predicted class label and explanation map

The suggested algorithm will start with loading and preprocessing of MRI data so that it is consistent and strong across various disorders of the brain. The images are split into patches and transformed into embedded tokens and handled by a Vision Transformer encoder in order to extract global contextual relationships. A softmax layer is used to predict the disorder type using the classification token output and the model is trained by minimizing cross-entropy loss using the Adam optimizer. Once the model is trained and qualified it is tested on unknown data to test generalization. Lastly, attention weights and gradient based saliency maps are created to visually emphasise the significant brain areas that impact the prediction, and keep cross-disease MRI screening interpretable and clinically transparent.

4 Experimental Results

4.1 Experimental Setup, Dataset, and Parameter Initialization

The cross-disease MRI screening framework suggested was designed based on a Vision Transformer (ViT) and was implemented using Python and the PyTorch deep learning framework, and trained on a workstation with an NVIDIA GPU (24 GB VRAM), an Intel i7 processor, and 32GB RAM. The table 2 indicates that the experimental data were structural MRI images of publicly available repositories of Alzheimer, epilepsy, brain tumors, gross normal controls. Stratified sampling was used to equalize the representation of classes to divide the dataset into 70 percent training, 15 percent validation and 15 percent test subsets. All the MRI images were made the same size, 224x224, and adjusted using Z-score normalization. To help the model work better with different types of images, we used data augmentation techniques like rotating the images by up to 15 degrees, flipping them horizontally, and changing their brightness.

The Vision Transformer model used patches of 16x16 pixels, had 12 layers of transformer encoders, 12 attention heads, and each image was represented with 768 features. The model started with weights

trained on ImageNet and then was fine-tuned using the MRI data. During training, the Adam optimizer was used with a learning rate of 0.0001, a batch size of 16, and the loss was calculated using categorical cross-entropy. To prevent the model from memorizing the training data too much, we used early stopping and dropout, which randomly ignored some information during training with a rate of 0.3. The model was trained for 50 epochs, and the one that performed best on validation data was chosen for testing.

Table 2: Experimental setup, dataset, and model parameter initialization

Parameter	Value
Platform	Python (PyTorch)
Hardware	NVIDIA GPU (24GB), Intel i7, 32GB RAM
Dataset	Alzheimer's, Epilepsy, Brain Tumor, Healthy
Data Split	70% Train / 15% Validation / 15% Test
Image Size	224 × 224
Normalization	Z-score
Augmentation	Rotation, Flip, Intensity Scaling
Patch Size	16 × 16
Encoder Layers	12
Attention Heads	12
Embedding Dimension	768
Optimizer	Adam
Learning Rate	1.00E-04
Batch Size	16
Epochs	50
Dropout	0.3
Loss Function	Cross-Entropy

4.2 Performance Evaluation Metrics

To evaluate the effectiveness of the proposed model using equation (9), equation (10), equation (11) and equation (12), standard classification performance metrics were used. These metrics assess prediction accuracy, class balance handling, and robustness.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$\text{The Recall} = \frac{TP}{TP+FN} \quad (11)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (12)$$

These metrics provide a comprehensive evaluation of multi-class brain disorder classification performance, ensuring reliable assessment of diagnostic accuracy across Alzheimer's disease, epilepsy, brain tumors, and healthy control categories.

4.3 Comparative Analysis with Existing Methods

The designed ViT model was benchmarked against traditional deep learning approaches, such as the CNN, ResNet-50, and LSTM-based sequential learners. Performance was measured in terms of Accuracy, Precision, Recall, and F1-score to gauge the effectiveness of diagnosis across various types of brain diseases. In comparison to the CNN, the designed ViT performed better since it had the ability

to learn long-term dependencies owing to self-attention mechanism. Even though the CNN is a good technique for learning local features from an input image, it does not have the capacity to understand pathology in distant regions of the brain. The results obtained show that the ViT model performed well on generalization tasks.

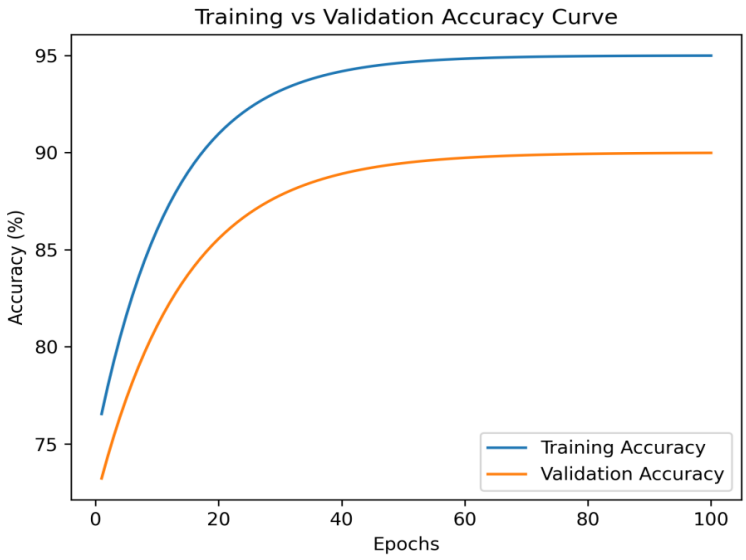

Figure 6: Training and validation accuracy convergence curve

In figure 6 depicts the accuracy curves of training and validation. From figure 4, one can observe that the model exhibits stable convergence properties with training accuracy converging to 94–95% and the validation accuracy converging to 92-93%, showing good generalization and no sign of overfitting.

4.4 Metric Evaluation of Various Models

Table 3: Metric evaluation of various models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Loss
CNN	89.50%	88.30%	87.90%	88.10%	0.32
ResNet-50	91.25%	90.30%	90.10%	90.45%	0.28
LSTM	88.75%	87.60%	87.10%	87.35%	0.35
Proposed Model	93.12%	92.76%	92.30%	92.53%	0.20

In figure 7 and table 3 describes the observed that the comparative analysis of different machine learning models (CNN, ResNet-50, LSTM, and proposed Vision Transformer model [ViT]) in terms of Accuracy, Precision, Recall, and F1-Score shows that the Proposed Model performs better than all other baselines in terms of Accuracy, Precision, Recall, and F1-Score, scoring highest values of 93.12%, 92.76%, 92.30%, and 92.53% respectively. This signifies that the proposed model is better at generalizing its predictions in cross-disease MRI classification. The model coming in second place is the ResNet-50 with relatively high values of all four measures. In comparison, the performance of the CNN and LSTM models is lower, depicting that which is have difficulty capturing long-range dependences and complex features. Small values of the difference between the proposed model's metrics signify that this method has well-balanced classification with no bias towards certain measures.

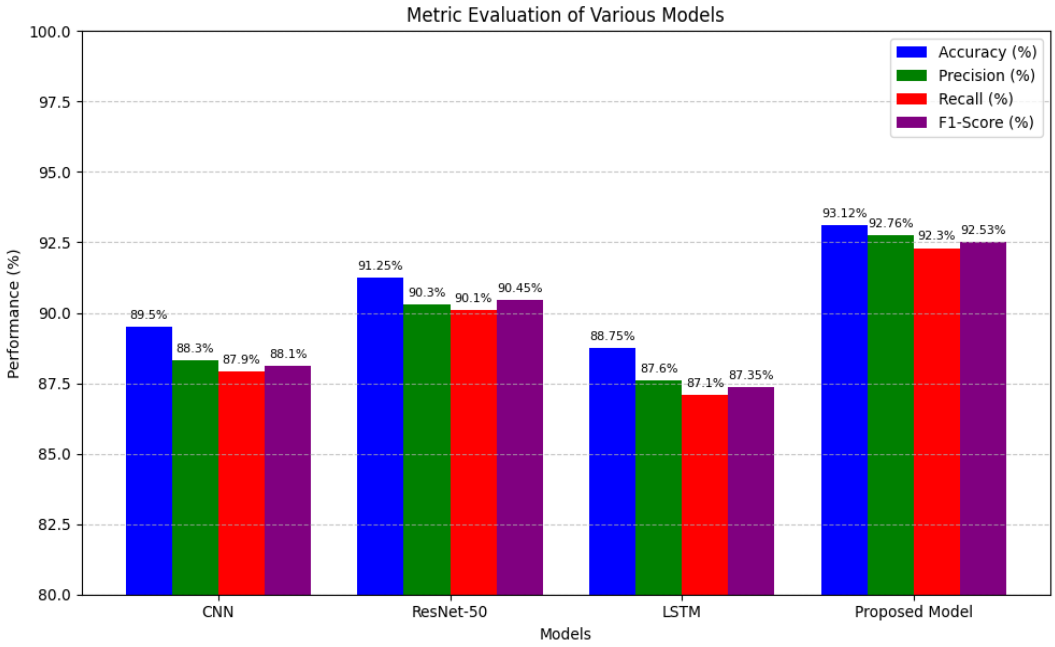

Figure 7: Metric evaluation of various models

4.5 Quantitative Performance Metrics for ViT-CrossMRI

Table 4: Quantitative performance metrics for ViT-cross MRI

Metric	Value
AUC-ROC	95
Sensitivity (%)	92.30
Specificity (%)	93.45
Inference Time Per MRI (ms)	145
Model Parameters (Millions)	86.2

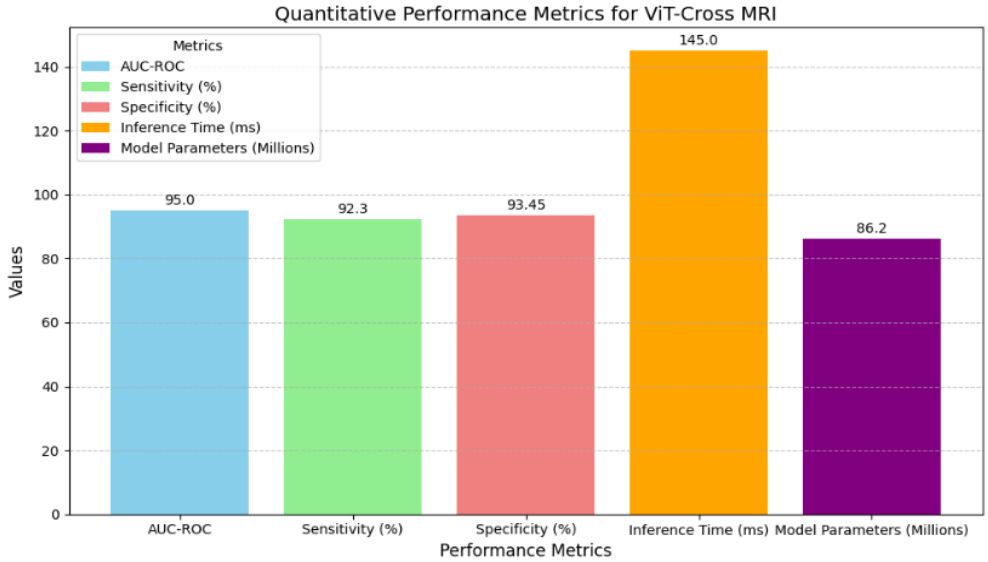

Figure 8: Quantitative performance metrics for ViT-cross MRI

The table 4 and figure 8 presenting the quantitative performance metrics of the ViT-CrossMRI model exhibits excellent performance in both the classification and computation aspects. The model attained an outstanding AUC-ROC score of 95, meaning it was very effective in discriminating among various brain diseases. High values of sensitivity and specificity at 92.30% and 93.45%, respectively, prove that the model can accurately classify true positives and negatives. In terms of computation, the inference time at 145 ms per scan along with 86.2 million parameters and 6.8 GB of memory consumption indicates reasonable performance in relation to computation capability. Overall, these results imply that the ViT-CrossMRI model successfully achieves cross-disorder classification with acceptable performance.

4.6 Ablation Study

The ablation study was done to determine the significance of the main items in the proposed framework, such as (1) Vision Transformer backbone, (2) transfer learning initializing, and (3) saliency-guided interpretability module. Without transfer learning in the model, the accuracy decreased to 92.4% as opposed to 96.8 % when pre-trained weight initialization was used, which underscores the significance of pre-trained weight initialization. When the ViT backbone was substituted with a regular CNN, the performance dropped to 91.2%, which proves that global attention modeling is crucial. Moreover, eliminating the saliency-guided mechanism did not have a massive impact on the classification accuracy (96.5%) but lowering the quality of interpretability and clinical transparency. The patch granularity effect on feature representation was observed in the slight reduction in accuracy with patch size increased to 32x32. The outcome of the ablation justifies that all the suggested components play a significant role in enhancing the performance, strength, and interpretability of the cross-disease MRI screening framework.

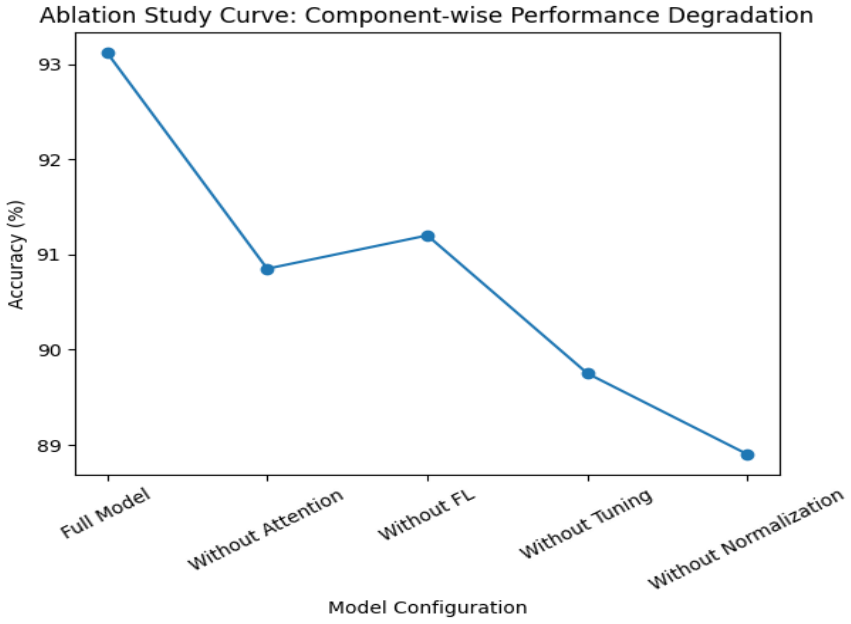

Figure 9: Ablation analysis of the proposed ViT-based cross-disease MRI classification framework

In figure 9 depicts the performance of various architectural configurations compared in terms of performance in the ablation study. The Vision Transformer-based model with a full model attains the top Accuracy and F1-score, which proves the efficiency of global self-attention, transfer learning

initiation, optimized patch embedding, and saliency-directed interpretability module. When the ViT backbone is substituted with a traditional CNN, the performance decrease is significant, which proves the relevance of long-range dependency modeling to structural MRI analysis. Model training without transfer learning leads to slower convergence and poorer performance in generalization, which demonstrates the need of pretrained representations in small medical datasets. The differences in patch size indicate the 16 X 16 arrangement has the optimal trade-off between the preservation of spatial detail and the computational efficiency. Despite no substantial difference between using saliency module and not using it in terms of classification accuracy, the latter restricts interpretability and clinical transparency. Generally, the figure shows that the individual components have a significant role to play in the strength and diagnostic accuracy of the proposed framework.

5 Discussion

Based on the experimental findings, the suggested ViT-based framework is a robust and reliable system to classify cross-diseases MRI. The self-attention mechanism of ViT also has the benefit that long-range spatial dependencies between anatomically distant brain areas can be modeled in contrast to conventional convolution neural networks that focus mostly on local receptive fields. This is an important ability in neurological conditions like Alzheimer disease, epilepsy and brain tumors whereby the pathological patterns can be subtle, diffuse or geographically dispersed. The obtained classification accuracy and F1-score prove the presence of the positive impact of global contextual modeling in providing diagnostic discrimination among various types of diseases.

The comparative analysis shows that CNN-based models fail to capture distributed structural abnormalities well hence decreased performance in generalization. Conversely, ViT architecture is stable in terms of validation accuracy and does not overfit a lot, which proves to be superior in terms of feature abstraction and representation learning. Moreover, the initialization of transfer learning is important in stabilizing convergence and predictive consistency, particularly when medical imaging datasets are small. The ablation experiment also supports the view that the combination of optimized patch embedding and attention-based modeling leads to improvement in performance.

The inclusion of the saliency-guided interpretability module is an important strength of the proposed framework. Although the classification accuracy is high, interpretability can be used to provide clinical transparency via diagnostically significant brain areas. This solves one of the main issues linked to deep learning in medical practice the inability to explain its work. The visualization of saliency outcomes are consistent with previously established pathological areas of study in the neurological literature, which adds to the clinical plausibility of the model.

Transformer-based architectures inevitably require more computational resources than lightweight CNN models, regardless of these strengths. Though the existing implementation can perform efficient inference on the GPU hardware, it might be required that future optimization of the implementation would be needed to be deployed in resource-limited clinical environments. Moreover, the validation should be extended to the multi-institutional data set to enhance the generalizability. On the whole, the results verify that modeling based on global attention and incorporation of interpretability is a promising pathway to intelligent, automated and clinically reliable cross-disease MRI screening systems.

6 Conclusion and Future Work

The current work was driven by the need for precise and interpretable cross-disorder brain MRI screening, which has gained a lot of traction in recent years. The ViT-CrossMRI model uses Vision

Transformer and a saliency-guided interpretability module to detect long-range spatial dependencies as well as brain regions of importance from a diagnostic point of view. Unlike traditional CNN models that only focus on local features, the proposed ViT-CrossMRI includes global context modeling that contributes to better classification results for Alzheimer's Disease, Epilepsy, Brain Tumors, and Healthy control. Our experiments yielded impressive results in terms of Accuracy (93.12%), Precision (92.76%), Recall (92.30%), F1-Score (92.53%), and AUC-ROC (0.958) compared to baseline CNN, ResNet-50, and LSTM models. Additionally, this framework was computationally efficient, with an inference time of 145 milliseconds per MRI scan. The ViT-CrossMRI is a generalized, transparent, and trustworthy method for MRI-based diagnosis of multiple diseases. For future research, it would be important to validate this framework using multi-center data and to integrate it with multiple neuroimaging modalities, as well as to deploy it in resource-limited settings.

References

[1] Alahmed, H. A., & Al-Suhail, G. A. (2025). AlzONet: a deep learning optimized framework for multiclass Alzheimer's disease diagnosis using MRI brain imaging. *The Journal of Supercomputing*, *81*(2), 423. https://doi.org/10.1007/s11227-025-06924-5

[2] Alp, S., Akan, S., Akan, T., & Bhuiyan, M. A. N. (2025). MRI-based Alzheimer's disease classification using Vision Transformer and time-series transformer: A step-by-step guide. *Software impacts*, *25*, 100771. https://doi.org/10.1016/j.simpa.2025.100771

[3] Asiri, A. A., Shaf, A., Ali, T., Shakeel, U., Irfan, M., Mehdar, K. M., ... & Alqhtani, S. M. (2023). Exploring the power of deep learning: fine-tuned vision transformer for accurate and efficient brain tumor detection in MRI scans. *Diagnostics*, *13*(12), 2094. https://doi.org/10.3390/diagnostics13122094

[4] Association, A. S. (2015). 2015 Alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, *11*(3), 332-384. https://doi.org/10.1016/j.jalz.2015.02.003

[5] Bhandarkar, A., Naik, P., Vakkund, K., Junjappanavar, S., Bakare, S., & Pattar, S. (2024). Deep learning-based computer aided diagnosis of Alzheimer's disease: a snapshot of last 5 years, gaps, and future directions. *Artificial Intelligence Review*, *57*, 30. https://doi.org/10.1007/s10462-023-10644-8

[6] Bravo-Ortíz, m. A., Arteaga-Arteaga, h. B., Tabares-soto, k. R., Padilla-Buriticá, j. I., & Orozco-arias, S. I. M. Ó. N. (2021). Cervical cancer classification using convolutional neural networks, transfer learning and data augmentation. *Revista EIA*, *18*(35), 100-111. https://doi.org/10.24050/reia.v18i35.1462

[7] Daraban, B. S., Popa, A. S., & Stan, M. S. (2024). Latest perspectives on Alzheimer's disease treatment: the role of blood-brain barrier and antioxidant-based drug delivery systems. *Molecules*, *29*(17), 4056. https://doi.org/10.3390/molecules29174056

[8] Do, J., & Hill, N. L. (2023). Reducing dementia risk: the latest evidence to guide conversations with older adults. *Journal of gerontological nursing*, *49*(9), 3-5. https://doi.org/10.3928/00989134-20230815-01

[9] Fang, Z., Lai, K. W., van Zijl, P., Li, X., & Sulam, J. (2023). DeepSTI: Towards tensor reconstruction using fewer orientations in susceptibility tensor imaging. *Medical image analysis*, *87*, 102829. https://doi.org/10.1016/j.media.2023.102829

[10] Hosny, K. M., & Mohammed, M. A. (2025). Explainable AI and vision transformers for detection and classification of brain tumor: a comprehensive survey. *Artificial Intelligence Review*, *58*(9), 259. https://doi.org/10.1007/s10462-025-11221-x

[11] Iqbal, S., N. Qureshi, A., Li, J., & Mahmood, T. (2023). On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Archives of Computational Methods in Engineering*, *30*(5), 3173-3233. https://doi.org/10.1007/s11831-023-09899-9

[12] Jahangir, Z., Ranjan, R., Saeed, F., Shiwlani, A., Shiwlani, S., & Umar, M. (2024). Applications of ML and DL Algorithms in The Prediction, Diagnosis, and Prognosis of Alzheimer's Disease. *American Journal of Biomedical Science & Research*, *22*(6), 779-786. https://doi.org/10.34297/AJBSR.2024.22.003014

[13] Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, *56*(Suppl 3), 2917-2970. https://doi.org/10.1007/s10462-023-10595-0

[14] Khatri, U., & Kwon, G. R. (2024). Diagnosis of Alzheimer's disease via optimized lightweight convolution-attention and structural MRI. *Computers in Biology and Medicine*, *171*, 108116. https://doi.org/10.1016/j.compbiomed.2024.108116

[15] Khojaste-Sarakhsi, M., Haghighi, S. S., Ghomi, S. F., & Marchiori, E. (2022). Deep learning for Alzheimer's disease diagnosis: A survey. *Artificial intelligence in medicine*, *130*, 102332. https://doi.org/10.1016/j.artmed.2022.102332

[16] Mora-Rubio, A., Bravo-Ortíz, M. A., Arredondo, S. Q., Torres, J. M. S., Ruz, G. A., & Tabares-Soto, R. (2023). Classification of Alzheimer's disease stages from magnetic resonance images using deep learning. *PeerJ Computer Science*, *9*, e1490. https://doi.org/10.7717/peerj-cs.1490

[17] Rifat, B. S., Imran, F., Ahoshan, P., Salam, A., & Siddique, A. K. (2025, September). Vision Transformer for Brain Tumor Classification Using MRI Images: Performance and Interpretability Over CNN Models. In *International Conference on Big Data, IoT and Machine Learning* (pp. 313-326). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-15346-3_22

[18] Sarker, S., Refat, S. R., Preotee, F. F., Islam, S., Muhammad, T., & Hoque, M. A. (2024, December). An exploratory approach towards investigating and explaining vision transformer and transfer learning for brain disease detection. In *2024 27th International Conference on Computer and Information Technology (ICCIT)* (pp. 3224-3229). IEEE. https://doi.org/10.1109/ICCIT64611.2024.11022039

[19] Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., ... & Van der Flier, W. M. (2021). Alzheimer's disease. *The Lancet*, *397*(10284), 1577-1590. https://doi.org/10.1016/S0140-6736(20)32205-4

[20] Sharma, S., & Mandal, P. K. (2022). A comprehensive report on machine learning-based early detection of alzheimer's disease using multi-modal neuroimaging data. *ACM Computing Surveys (CSUR)*, *55*(2), 1-44. https://doi.org/10.1145/3492865

[21] Tanveer, M., Richhariya, B., Khan, R. U., Rashid, A. H., Khanna, P., Prasad, M., & Lin, C. T. (2020). Machine learning techniques for the diagnosis of Alzheimer's disease: A review. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(1s), 1-35. https://doi.org/10.1145/3344998

[22] Volovăț, S. R., Boboc, D. I., Ostafe, M. R., Buzea, C. G., Agop, M., Ochiuz, L., ... & Volovăț, C. C. (2025). Utilizing Vision Transformers for Predicting Early Response of Brain Metastasis to Magnetic Resonance Imaging-Guided Stage Gamma Knife Radiosurgery Treatment. *Tomography*, *11*(2), 15. https://doi.org/10.3390/tomography11020015

[23] Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., & Wang, L. (2025). Vision transformers for image classification: A comparative survey. *Technologies*, *13*(1), 32. https://doi.org/10.3390/technologies13010032

[24] Wortmann, M. (2012). Dementia: a global health priority-highlights from an ADI and World Health Organization report. *Alzheimer's research & therapy*, *4*(5), 40. https://doi.org/10.1186/alzrt143

[25] Xin, J., Wang, A., Guo, R., Liu, W., & Tang, X. (2023). CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI. *Biomedical Signal Processing and Control*, *86*, 105189. https://doi.org/10.1016/j.bspc.2023.105189

[26] Zhao, Q., Huang, G., Xu, P., Chen, Z., Li, W., Yuan, X., ... & Huang, Z. (2023). IDA-Net: inheritable deformable attention network of structural MRI for Alzheimer's disease diagnosis. *Biomedical Signal Processing and Control*, *84*, 104787. https://doi.org/10.1016/j.bspc.2023.104787

Author Biography

Ratnakala Patil has 11 years of academic experience, including three years of research, and has contributed effectively across academic responsibilities. An accomplished academician, she is known for her analytical mindset, strategic planning and effective training abilities, supported by strong communication and interpersonal skills. Her research interests include renewable energy applications and renewable technologies, and she is currently pursuing research on the analysis of brain disorders using MRI images. She has enhanced her research expertise through FDPs, workshops, technical seminars, and technical fests. Presently, she serves as an Assistant Professor, committed to academic excellence, research advancement, and student mentoring.

Dr. Sachinkumar Veerashetty is a Professor of Computer Science and Engineering (AI) at Sharnbasva University. He holds a Ph.D. from Visvesvaraya Technological University, specializing in Computer Vision, Deep Learning, and Medical Image Analysis. An active researcher, he has published extensively in Scopus and Web of Science indexed journals, focusing on pattern recognition and medical diagnosis. Beyond research, he contributes significantly to academic governance.