ISSN 2723-6471

Transformer Architectures for Automated Brain Stroke Screening from MRI Images

Husni Teja Sukmana[1,*], Zainal Arifin Hasibuan[2], Abdul Wahab Abdul Rahman[3] Luhur Bayuaji[4], Siti Ummi Masruroh[5]

[1,5]Department of Informatic, Faculty of Science Technology, State Islamic University Syarif Hidayatullah Jakarta, 15412, Indonesia

[2]Faculty of Computer Science, Dian Nuswantoro University, Indonesia

[3]Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Malaysia

[4]Faculty of Data Science and Information Technology (FDSIT), INTI International University, Nilai, Malaysia

(Received: January 4, 2025; Revised: February 25, 2025; Accepted: April 11, 2025; Available online: May 14, 2025)

Abstract

Early and accurate detection of stroke is critical for timely intervention and improved patient outcomes. This study evaluates the effectiveness of deep learning models—specifically the Vision Transformer (ViT)—in the automated classification of brain stroke conditions using MRI images. A curated dataset of brain MRI scans was used to train and assess three models: a convolutional neural network baseline (ResNet-18), a standard ViT (ViT-Base), and a modified ViT architecture. All models were fine-tuned using transfer learning under consistent preprocessing, training configurations, and evaluation protocols to ensure a fair comparison. The experimental results show that the modified ViT achieved the best overall performance, with a validation accuracy of 97.2% and an AUC score of 0.9934, outperforming both ResNet-18 (92.6% accuracy, AUC 0.9982) and ViT-Base (78.6% accuracy, AUC 0.8278). In particular, the modified ViT demonstrated stronger class-wise precision, recall, and F1-scores in detecting stroke cases, suggesting enhanced sensitivity to critical image features. ResNet-18 remained competitive and efficient, especially in stroke recall (0.99), while ViT-Base underperformed—highlighting the need for architectural and training adaptations when applying ViT models to small medical datasets. Visualization techniques such as confusion matrices, ROC curves, integrated gradients, and attention maps further validated the robustness and interpretability of the modified ViT model. However, the ViT-based models required significantly higher computational resources and longer training times compared to the CNN baseline, which may limit their deployment in real-time or resource-constrained clinical settings. These findings support the growing potential of transformer-based architectures in medical image classification, particularly for stroke diagnosis, provided that computational costs are weighed against diagnostic benefits.

Keywords: Vision Transformer, Stroke Diagnosis, Medical Imaging, Deep Learning, Convolutional Neural Network, Image Classification, Machine Learning

1. Introduction

Stroke remains one of the leading causes of mortality and long-term disability worldwide. According to the World Health Organization, millions of people suffer strokes each year, with many experiencing irreversible neurological damage or death [1]. Prompt and accurate diagnosis is essential, as early identification of stroke type, location, and severity directly guides time-sensitive therapeutic interventions such as thrombolysis or surgery [2], [3]. Clinical expectations for diagnostic accuracy in stroke imaging generally require high sensitivity and specificity—often exceeding 85% and 90%, respectively—when compared to expert radiologist interpretation [4]. Conventional stroke diagnosis relies heavily on radiological imaging, particularly Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), interpreted by medical specialists. However, this process is subject to delays, human error, and inter-observer variability [5], [6].

In response to these challenges, machine learning (ML) has emerged as a promising solution to enhance diagnostic precision and efficiency in medical imaging. ML models, particularly those based on supervised learning, can be trained

***Corresponding author: Husni Teja Sukmana (husniteja@uinjkt.ac.id)**

DOI: https://doi.org/10.47738/jads.v6i2.736

to identify patterns and anomalies in complex imaging data, offering support in classification and decision-making tasks [7]. Among ML approaches, deep learning (DL)—a subset of ML—has shown exceptional performance in medical image analysis, with Convolutional Neural Networks (CNNs) widely used for tasks such as classification, detection, and segmentation [8], [9]. These models excel at extracting hierarchical features but are limited by their localized receptive fields, which restrict their ability to capture the global context—an important aspect for interpreting complex brain pathologies such as stroke [10]. While techniques like dilated convolutions and global average pooling have been introduced to mitigate this limitation, they offer only partial solutions [11], [12].

To address these constraints, Transformer-based architectures—originally developed for Natural Language Processing—have been adapted to computer vision tasks [13]. The Vision Transformer (ViT), in particular, processes images as sequences of patches and leverages self-attention mechanisms to model long-range spatial relationships across the entire image [14], [15]. This capability is especially advantageous in medical contexts where lesions may be subtle, diffuse, or distributed across multiple brain regions. Recent studies have shown ViT's effectiveness in various medical domains, including dermatology [16], ophthalmology [17], and increasingly, neuroimaging [18].

This study explores the application of Vision Transformers for the classification of stroke from brain MRI scans. A ViT model is fine-tuned on a curated stroke dataset and benchmarked against a conventional CNN baseline, ResNet18, under identical preprocessing and training conditions. The objective is to assess ViT's classification performance, analyze its advantages and limitations in clinical stroke diagnosis, and evaluate its feasibility for real-world deployment [19]. By investigating the role of Transformer-based models in brain stroke classification, this research contributes to the growing body of literature on machine learning applications in medical diagnostics and aims to support the development of faster, more accurate, and scalable solutions for clinical decision-making [20].

2. Literature Review

2.1. Stroke Diagnosis and Medical Imaging

Stroke diagnosis fundamentally relies on neuroimaging modalities such as Magnetic Resonance Imaging (MRI) and CT, which enable the identification of ischemic or hemorrhagic events, as well as the localization and extent of tissue damage. These imaging techniques are critical for informing timely therapeutic interventions. However, image interpretation depends heavily on the radiologist's expertise and is subject to variability and diagnostic delays, particularly under time-critical conditions [21]. Given the growing demand for rapid and standardized diagnostics and the global shortage of radiology specialists, artificial intelligence (AI) has emerged as a powerful tool to assist in automated stroke detection [22].

Earlier applications of AI in stroke diagnosis primarily relied on conventional machine learning algorithms, such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forests [23]. These systems often required labor-intensive preprocessing and hand-crafted feature engineering—limiting their scalability and performance across diverse datasets and imaging protocols.

2.2. Deep Learning in Medical Image Analysis

The advent of deep learning has revolutionized the field of medical imaging by enabling automated feature learning and end-to-end model training. CNNs, in particular, have achieved remarkable success in clinical tasks including tumor segmentation, disease classification, and lesion detection [24]. In stroke imaging, CNNs have been successfully applied for tasks such as infarct region segmentation, stroke subtype classification, and lesion volume estimation [25].

Among CNN architectures, ResNet stands out due to its use of residual connections, which address the vanishing gradient problem and allow for deeper, more expressive networks [26]. Its simplicity and effectiveness make it a popular baseline in medical imaging research. However, CNNs are inherently constrained by their local receptive fields, limiting their ability to capture global spatial relationships—a significant drawback in neuroimaging, where lesions may be diffuse or distributed across multiple brain regions [27].

2.3. Vision Transformers in Medical Imaging

To address these limitations, ViT, adapted from the Transformer architecture originally designed for natural language processing, have been introduced to vision tasks [28]. ViT treats images as sequences of non-overlapping patches and employs multi-head self-attention to model complex and long-range dependencies across the entire image. This architectural shift allows ViT to learn global contextual representations, offering a distinct advantage over CNNs in tasks requiring holistic understanding.

In medical imaging, ViT has shown promise in several applications such as skin cancer detection, diabetic retinopathy screening, and chest X-ray classification [29]. In the domain of neuroimaging, ViT is emerging as a competitive alternative to CNNs. For instance, Abbaoui et al. applied a ViT model for ischemic stroke classification on MRI scans and reported higher accuracy compared to standard CNN architectures, attributing the improvement to ViT's global attention mechanism [30]. Similarly, Hossain et al. introduced a hybrid ViT-LSTM architecture for stroke detection from CT images, demonstrating that the self-attention mechanism enhances the detection of subtle stroke manifestations that CNNs may overlook [31].

Nonetheless, the use of ViT in medical imaging poses challenges. Unlike CNNs, which incorporate inductive biases such as locality and translation invariance, ViT requires larger training datasets to generalize effectively. This makes it more prone to overfitting, especially in the context of small-scale medical datasets. Thus, successful ViT applications often necessitate pretraining on large datasets, extensive data augmentation, or the integration of hybrid architectural strategies [32].

While CNNs remain a strong choice for many medical imaging tasks due to their efficiency and architectural maturity, Vision Transformers provide a powerful alternative for cases requiring a deeper understanding of global image structure. Their application in stroke diagnosis is promising and warrants further exploration, particularly as more data-efficient and explainable transformer models continue to emerge.

3. Methodology

This section outlines the experimental framework adopted for evaluating the performance of ViT and ResNet18 models in brain stroke classification. The methodology includes data preparation, preprocessing and augmentation, model configuration, training pipeline, and performance evaluation. The end-to-end research pipeline is depicted in figure 1, which summarizes the major stages from data preparation to interpretability analysis and final discussion. The workflow ensures consistency across experiments by applying identical preprocessing and training configurations to both architectures.

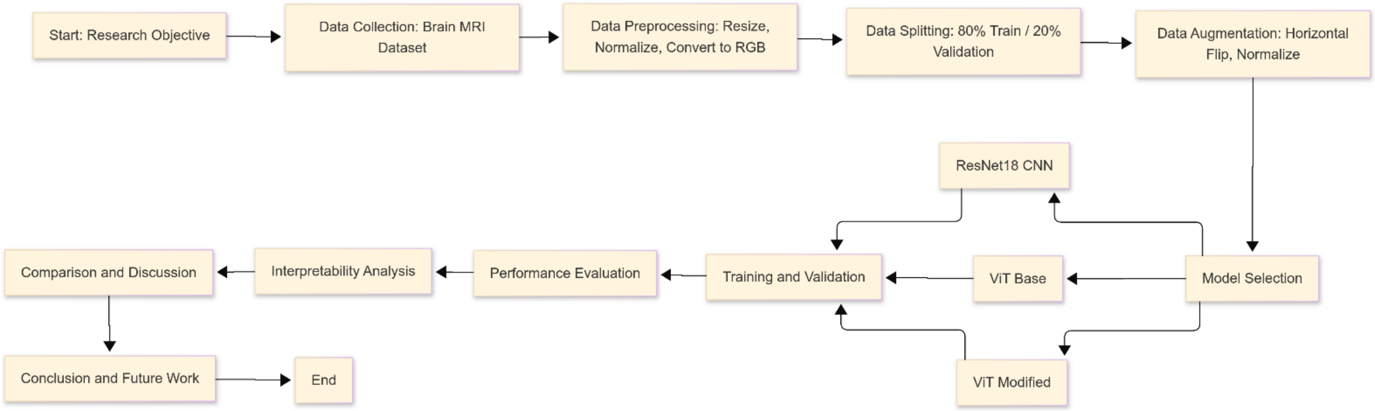
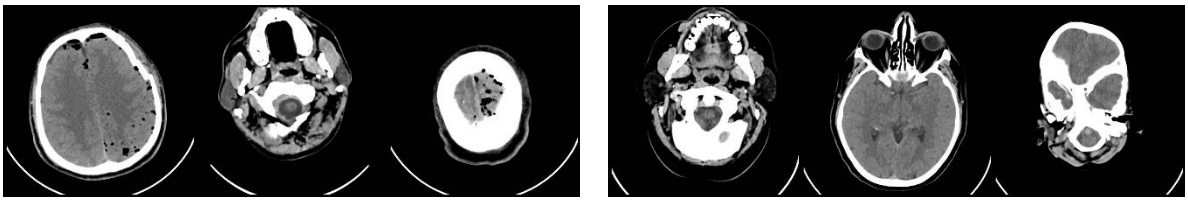

Start: Research Objective
Data Collection: Brain MRI Dataset
Data Preprocessing: Resize, Normalize, Convert to RGB
Data Splitting: 80% Train / 20% Validation
Data Augmentation: Horizontal Flip, Normalize
ResNet18 CNN
Model Selection
ViT Base
ViT Modified
Training and Validation
Performance Evaluation
Interpretability Analysis
Comparison and Discussion
Conclusion and Future Work
End

Figure 1. General research workflow for model comparison and evaluation

The research begins by defining a clear objective, specifically focused on classifying brain MRI images using deep learning models. A publicly available brain MRI dataset is collected and subjected to a series of preprocessing steps, including resizing, normalization, and RGB conversion to ensure compatibility with model input requirements. The preprocessed data is then split into 80% training and 20% validation sets. To enhance model generalization and address overfitting, data augmentation techniques such as horizontal flipping and further normalization are applied. Following this, three models are selected for evaluation: ResNet18 CNN, ViT Base, and a modified version of the ViT. Each model undergoes training and validation using the augmented dataset. Model performance is subsequently assessed using standard evaluation metrics, and an interpretability analysis is conducted to gain insights into the decision-making process of each model. The results are compared and discussed to highlight the strengths and limitations of each architecture. Finally, the study concludes with a summary of findings and outlines directions for future work.

3.1. Dataset Description and Preprocessing

The dataset used in this study comprises 2,501 brain MRI scans obtained from a previously published research study [33], each labeled as either stroke or normal. To simplify the classification task and better reflect real-world stroke screening scenarios, both ischemic and hemorrhagic stroke subtypes were grouped under a single "stroke" category, resulting in a binary classification problem. All images were preprocessed by resizing to 224 × 224 pixels and normalized using a mean and standard deviation of [0.5, 0.5, 0.5]. Although the original MRI scans are in grayscale, they were converted to 3-channel RGB format to align with the input requirements of pretrained models. The dataset was split into 80% for training and 20% for validation using a fixed random seed to ensure reproducibility. Basic data augmentation techniques—such as random horizontal flipping—were applied during training to enhance model generalization. No color jittering or affine transformations were used, as they do not represent realistic variability in medical imaging. Figure 2 shows a sample dataset used.



Normal

Stroke

Figure 2. Sample Dataset

3.2. Model Architectures

3.2.1. ResNet18

ResNet18 was employed as the baseline CNN model in this study. The architecture comprises 18 layers organized into 4 sequential residual blocks, which incorporate shortcut connections to facilitate efficient gradient flow and mitigate the vanishing gradient problem during backpropagation. To adapt the model for the binary classification task, the original fully connected layer was replaced with a custom classification head consisting of a single output neuron with a sigmoid activation function. The model was initialized with pretrained weights from ImageNet and subsequently fine-tuned on the target dataset to enhance task-specific performance. A visual overview of ResNet18's role within the overall processing pipeline is presented in figure 1. For interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized to generate visual explanations of the model's predictions, highlighting the most informative regions in the brain MRI images that contributed to the final classification decisions.

3.2.2. ViT

Two ViT models were explored in this study. ViT Base refers to the standard ViT-Base architecture, which utilizes patch embeddings, 12 attention heads, and 12 transformer encoder layers. A classification (CLS) token is appended to the input sequence and propagated through the encoder stack, with its final output used for classification. This model is fully fine-tuned using ImageNet-pretrained weights. ViT Modified is an enhanced variant of the base model, optimized by adjusting learning rates and incorporating interpretability tools such as LIME to improve model explainability. The internal structure of the ViT model is depicted in figures 3 and figure 4, illustrating the process from image patch embedding through multi-head self-attention, culminating in classification via the CLS token output.

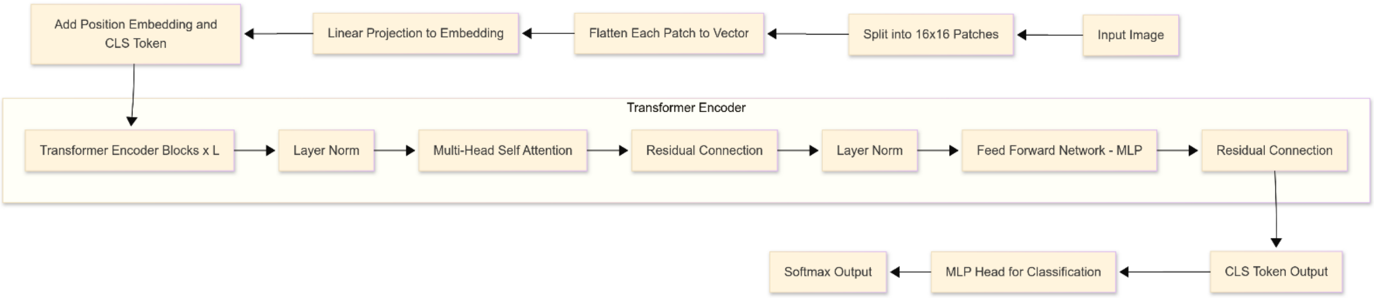

Add Position Embedding and CLS Token
Linear Projection to Embedding
Flatten Each Patch to Vector
Split into 16x16 Patches
Input Image
Transformer Encoder
Transformer Encoder Blocks x L
Layer Norm
Multi-Head Self Attention
Residual Connection
Layer Norm
Feed Forward Network - MLP
Residual Connection
Softmax Output
MLP Head for Classification
CLS Token Output

Figure 3. Vision Transformer architecture with transformer encoder and classification output

The ViT architecture processes input images by first dividing them into fixed-size patches, typically 16×16 pixels. Each patch is then flattened into a one-dimensional vector and linearly projected into an embedding space. A special CLS token is prepended to the sequence of patch embeddings, and positional embeddings are added to retain spatial information. This sequence is fed into a series of Transformer encoder blocks, each comprising layer normalization, multi-head self-attention mechanisms, residual connections, and a feed-forward multilayer perceptron (MLP). These components enable the model to capture global dependencies across all patches. After passing through all encoder layers, the output corresponding to the CLS token is extracted, serving as a compact representation of the entire image. This token is then passed through an MLP head for classification, followed by a softmax layer to produce the final class probabilities. The architecture leverages self-attention to model complex spatial relationships without relying on convolutional operations, offering an alternative paradigm for image classification tasks.

The ViT Modified architecture (see figure 4) developed in this study extends the standard ViT by introducing several key enhancements tailored to the characteristics of brain MRI classification tasks. The model pipeline begins with an input image of size 224×224 pixels, which is first converted to RGB format to ensure compatibility with pretrained weight structures and consistent channel dimensions. The image is then split into a fixed number of non-overlapping square patches (typically 16×16 pixels), effectively transforming the spatial domain into a sequence of image tokens.

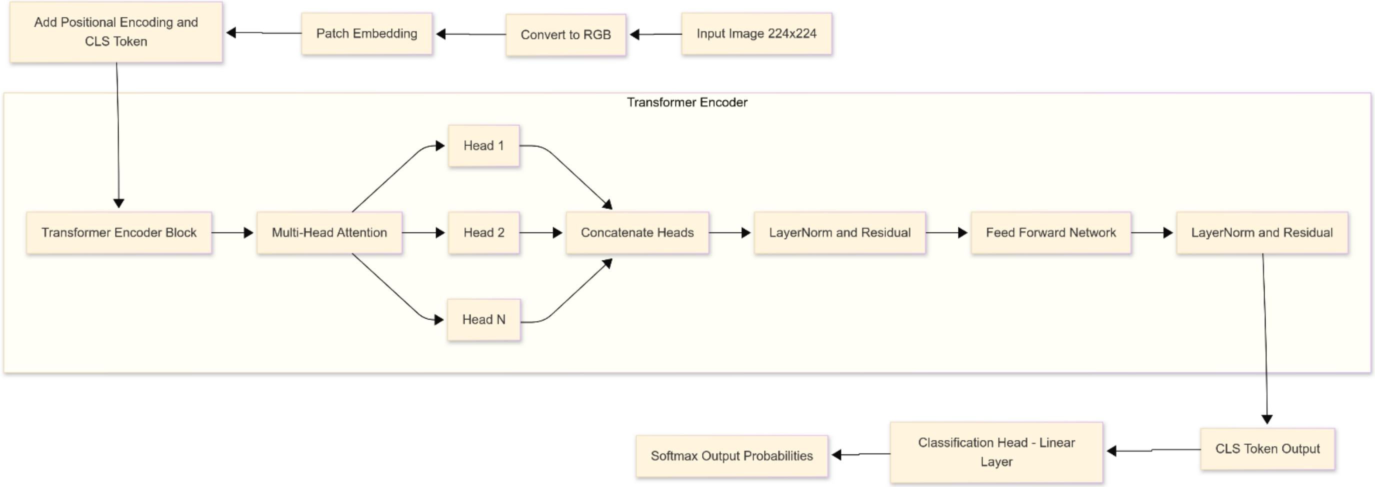


Figure 4. Patch embedding and encoder design in ViT-Base

Each patch is flattened and linearly projected into a high-dimensional embedding space. To this sequence of patch embeddings, a CLS token is prepended. This special token is intended to aggregate global information across the image during attention processing and serves as the final representation used for classification. Additionally, positional encodings are added to each token in the sequence, enabling the model to retain spatial order — a critical feature in medical image analysis where location context matters.

The encoded sequence is then passed into a custom Transformer Encoder block, which constitutes the core innovation of the ViT Modified design. Within this block, a Multi-Head Self-Attention (MHSA) mechanism is employed, wherein the input is processed through N independent attention heads. Each head captures different relational patterns between patches, allowing the model to attend to multiple spatial contexts simultaneously. The outputs from these parallel attention heads are concatenated and combined into a unified representation.

This attention output undergoes Layer Normalization and residual connections, both before and after passing through a Feed Forward Network (FFN) composed of fully connected layers with non-linear activation (e.g., GELU or ReLU). These architectural components are repeated in multiple stacked encoder layers (not explicitly shown in the simplified diagram) to allow the model to learn progressively abstract representations. The residual pathways and normalization layers are critical for stabilizing training and ensuring effective gradient flow across deep layers.

Once the sequence has passed through the full encoder stack, the updated CLS token output is extracted and routed to a classification head, which consists of a linear (dense) layer that maps the high-dimensional CLS embedding to class logits. These logits are then passed through a softmax function to produce output probabilities for the target classes. In this study, the classification task is binary, so the final layer is configured accordingly.

This modified architecture provides improved interpretability, modular attention control, and better adaptation to domain-specific data like medical imaging. Moreover, it is designed to be compatible with explainability tools such as LIME or Attention Rollout, which can be applied directly to the attention weights or CLS token behavior to visualize decision rationales. The detailed flow of this architecture is illustrated in figure 4.

Algorithm 1 explains the process of image classification using a modified Vision Transformer architecture. The input image is first converted to RGB and divided into fixed-size patches. Each patch is flattened and projected into an embedding space. A learnable [CLS] token is added to the sequence, followed by the addition of positional encoding. The sequence is then passed through multiple transformer encoder layers, where self-attention and feed-forward operations are applied. Finally, the representation of the [CLS] token is used for classification through a linear layer followed by softmax to produce the output class probabilities.

Algorithm 1. ViT_Modified — Vision Transformer with Modified Pipeline

```
Input:
$I \in \mathbb{R}^{H \times W \times C}$: Input Image
$Patch\ size\ P = 16$
Embedding dimension $D$, Number of layers $L$, Number of heads $H$, Number of classes $K$
Output:
$p \in \mathbb{R}^{K}$: Class probabilites

1:  Ensure Input Validity
    $if\ C \neq 3\ then\ I \leftarrow convert\_to\_RGB(I)$
2:  Divide Image into Patches
    $N = \frac{H \times W}{P^2}, X = reshape(I, (N, P^2 \cdot C)) \in \mathbb{R}^{N \times (P^2 \cdot C)}$
3:  Linear Projection to Patch Embedding
    $Z = X \cdot W_e + b_e \in \mathbb{R}^{N \times D}$
4:  Add [CLS] Token
    $Z_{cls} \in \mathbb{R}^{1 \times D}, Z = concat(Z_{cls}, Z) \in \mathbb{R}^{(N+1) \times D}$
5:  Add Positional Encoding
    $P_E = position_{encoding(N+1,D)},\quad Z^{II} = Z^{I} + P_E$
6:  Transformer Encoder Layers
    For $l = 1$ to $L$ do
        Multi-head Attention:
        For each head $h = 1$ to $H$:
            $Q_h = Z^{II} \cdot W_Q^h,\ K_h = Z^{II} \cdot W_K^h,\ K_h = Z^{II} \cdot W_V^h$
            $Attention_h = softmax(\frac{Q_h \cdot K_h^T}{\sqrt{d_k}}) \cdot V_h$
        End for
        $MHA = concat(Attention_1, \dots, Attention_H)$
        $Z_{res1} = LayerNorm(Z^{II} + MHA)$
        Feed Forward Network:
        $FFN(Z_{res1}) = GELU(Z_{res1}, W_1, b_1) \cdot W_2 + b_2$
        $Z^{II} = LayerNorm(Z_{res1} + FFN(Z_{res1}))$
    End For
7:  Extract [CLS] Token Representation
    $Z_{final} = Z^{II}[0] \in \mathbb{R}^{D}$
8:  Classification Head
    $y = Z_{final} \cdot W_{cls} + b_{cls} \in \mathbb{R}^{K}$
9:  Softmax Output
    $p_i = \frac{e^{y_i}}{\sum_{j=1}^{K} e^{y_j}},\quad \forall_i = 1 \dots K$
10: Return Prediction
    $return\ p$
```

3.3. Training Configuration

Both ResNet18 and ViT models were trained under identical conditions for fairness. Cross-Entropy Loss was used as the objective function. For optimization, AdamW was used for ViT models, while ResNet18 used the Adam optimizer.

The learning rate was initialized at 3 × 10^{-5} for both, with a StepLR scheduler reducing it by a factor of 0.1 after 5 epochs. Training was conducted for 10 epochs with a batch size of 32 using GPU acceleration (NVIDIA RTX 4050). Validation was performed at the end of each epoch, and the best-performing model based on validation accuracy was saved for evaluation.

3.4. Evaluation Metrics

Model performance was comprehensively assessed using several evaluation metrics. Accuracy was calculated as the proportion of correctly predicted samples relative to the total samples:

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \tag{1}$$

Precision measured the ratio of true positive predictions to all positive predictions:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (or sensitivity) quantified the proportion of true positive cases identified:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1-Score, representing the harmonic mean of precision and recall, was computed as:

$$F1 = 2 \times \frac{(Precision \times Recall)}{Precision + Recall} \tag{4}$$

Additional robust metrics were also evaluated. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed to assess the model's ability to distinguish between classes across various threshold settings. Cohen's Kappa coefficient was calculated to measure inter-rater agreement beyond chance level. Confusion matrices were generated to visualize the distributions of true positives, true negatives, false positives, and false negatives. Additionally, training and validation loss curves were plotted over epochs to inspect the models' convergence behavior and to monitor potential overfitting

4. Results and Discussion

This section presents the experimental results comparing the performance of ViT architectures—both base and modified—with a baseline CNN, ResNet-18, for the classification of brain stroke images using MRI data. All models were trained and evaluated under identical conditions to ensure a fair comparison. Key aspects evaluated include training dynamics, classification metrics, computational efficiency, and qualitative outputs.

4.1. Training and Validation Accuracy

Training and validation accuracies were monitored across five epochs for all models. Table 1 presents the validation accuracies for ViT-Base, ViT-Modified, and ResNet-18 during the training process. Several trends can be observed from the results. The ViT-Modified model exhibited rapid convergence, achieving over 92% validation accuracy after just one epoch and peaking at 97.2% by epoch 4. Its early and stable improvement suggests that the model effectively captured discriminative features from stroke MRI scans with minimal training. The consistent performance over the final epochs also indicates convergence and strong generalization.

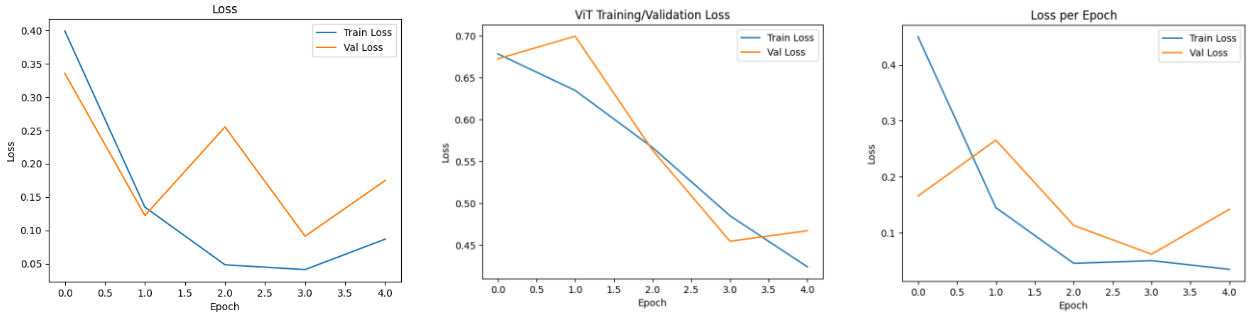
Table 1. Validation Accuracy (%) Over 5 Epochs

Epoch	ViT-Base	ViT-Modified	ResNet-18
1	60.9	92.8	83.4
2	63.1	90.2	95.4
3	73.5	96.4	91.4
4	79.2	97.2	96.4
5	78.6	96.4	92.6

ResNet-18 showed steady improvement, reaching its best validation accuracy of 92.6% at epoch 5. Although slightly less accurate than ViT-Modified, it remained a strong and efficient performer, especially in early stroke detection, as evidenced by its high recall scores. In contrast, ViT-Base demonstrated the slowest learning progression, with lower validation accuracy throughout. This underperformance may be attributed to its sensitivity to training hyperparameters and limited inductive bias, which makes it less effective on small datasets without further architectural or optimization refinements. Overall, these results indicate that transformer-based models, particularly the modified ViT, offer superior learning efficiency and classification accuracy compared to traditional CNNs. Their ability to model global dependencies and subtle spatial patterns contributes to better generalization in complex medical imaging tasks.

4.2. Training Loss Convergence

To analyze the learning efficiency and convergence behavior of each model, the training loss was recorded over five epochs and visualized in figure 5. The loss curves provide quantitative insights into how effectively each model optimized its parameters in response to the training data. At the outset, all three models began with relatively high loss values, as expected from uninitialized weights and early exposure to the data. However, the modified Vision Transformer (ViT-Modified) demonstrated the most rapid decline in loss. Starting from approximately 0.45 in the first epoch, its loss dropped sharply to around 0.045 by the third epoch and stabilized between 0.03 and 0.04 by the fourth and fifth epochs. This early and steady flattening of the curve suggests that the model quickly learned discriminative features and reached a state of convergence with minimal further optimization required.



Resnet-18 ViT (base) ViT (modified)

Figure 5. Training loss

In contrast, the ResNet-18 model exhibited a more gradual and linear reduction in loss. Beginning at around 0.60, its training loss decreased steadily and reached approximately 0.087 by the fifth epoch. Although this trajectory reflects consistent learning, the slower rate of convergence indicates that the CNN architecture required more iterations to achieve similar levels of optimization as the transformer-based model. Meanwhile, the baseline ViT model (ViT-Base) showed the weakest convergence. Despite some decrease in training loss, it plateaued at a much higher value of around 0.42, indicating difficulties in fitting the data effectively. This performance gap likely stems from the lack of architectural regularization and the known sensitivity of ViT models to smaller datasets when not adequately tuned or pre-trained.

These distinct convergence patterns highlight the relative strengths and limitations of each architecture. The ViT-Modified model benefited from a carefully chosen optimization strategy—employing the AdamW optimizer with a small learning rate and appropriate weight decay—which likely contributed to its smooth and rapid descent. Additionally, the self-attention mechanism intrinsic to the transformer architecture allowed the model to capture global contextual relationships between image patches early in training, which is particularly advantageous in analyzing complex medical images such as brain MRI scans. In contrast, the ResNet-18 model, while inherently stable and computationally efficient, relies more heavily on local receptive fields and hierarchical feature extraction, which may explain its slower but reliable optimization behavior.

Although low training loss generally indicates strong learning performance, it is not a definitive measure of generalization. Particularly in the context of limited datasets, low training loss can signal overfitting if not supported by equally strong validation performance. However, in this study, the validation accuracy of the ViT-Modified model remained closely aligned with its training loss trajectory, suggesting that it was not merely memorizing training data but was instead learning meaningful and generalizable representations. This conclusion is further supported by interpretability tools such as integrated gradients and attention maps, which confirmed the model's focus on clinically relevant image regions.

In summary, the training loss analysis demonstrates that the ViT-Modified model not only converges faster and more smoothly than its counterparts but also achieves superior optimization with fewer iterations. These characteristics, when combined with strong validation performance and model interpretability, underline the effectiveness of the transformer-based approach in complex medical image classification tasks.

4.3. Confusion Matrix Analysis

To gain deeper insight into the classification behavior of each model, confusion matrices were generated for ResNet-18, ViT-Base, and ViT-Modified on the validation set. These matrices summarize the models' predictions relative to the ground truth and reveal the types of classification errors each architecture tends to make. Figure 6 displays the confusion matrix of ResNet-18. The model correctly identified 269 out of 305 "Normal" cases and 195 out of 196 "Stroke" cases. It misclassified 36 normal cases as stroke (false positives) and only 1 stroke case as normal (false negative). This corresponds to a stroke recall (sensitivity) of 99.5% and a normal recall (specificity) of approximately 88.2%. Although the model exhibits a slight tendency toward false positives—erroneously labeling healthy patients as having stroke—it maintains high recall for stroke, which is particularly valuable in clinical settings where missed diagnoses carry serious risks.

Figure 7 shows the confusion matrix for the ViT-Base model. While it correctly classified 274 out of 305 "Normal" cases, its performance on "Stroke" cases was substantially weaker: only 120 out of 196 stroke instances were correctly predicted, resulting in 76 false negatives. This yields a stroke recall of just 61.2%, a considerable drop compared to ResNet-18. Despite having a slightly higher specificity for normal cases (89.8%), the large number of missed stroke cases is clinically problematic. The model's poor recall indicates an inability to reliably capture features characteristic of stroke, which may be due to insufficient optimization or the lack of architectural regularization in the base transformer model. In contrast, figure 8 presents the confusion matrix for the ViT-Modified model, which clearly outperforms both ViT-Base and ResNet-18. It correctly classified 307 out of 311 "Normal" cases and 176 out of 190 "Stroke" cases, resulting in 14 false negatives and just 4 false positives. This corresponds to a stroke recall of 92.6% and a normal recall of 98.7%. The balance between high sensitivity and specificity makes the ViT-Modified both reliable and clinically robust. Unlike the base model, the modified architecture is able to detect stroke features with higher accuracy and consistency across cases.

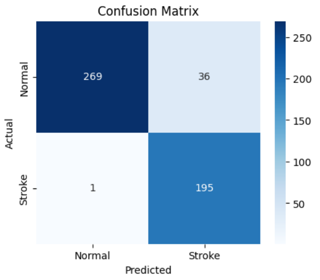


Figure 6. Resnet-18 Confusion Matrix

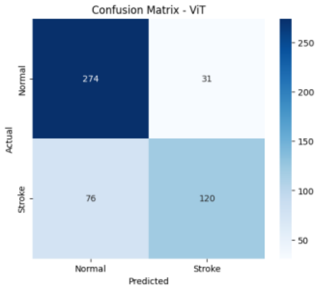


Figure 7. ViT (base) Confusion Matrix

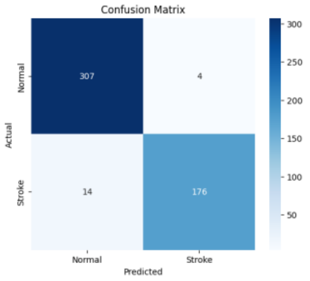


Figure 8. ViT (modified) Confusion Matrix

Overall, the confusion matrices reveal that the ViT-Modified model achieves the best trade-off between detecting stroke cases accurately and avoiding false alarms. ResNet-18, while slightly less precise, remains a strong alternative, especially considering its computational efficiency. On the other hand, ViT-Base underperforms significantly, particularly in stroke recall, making it less suitable for critical diagnostic applications without further tuning or architectural adjustments. The superior performance of ViT-Modified can be attributed to its global attention mechanism, which allows it to model spatial dependencies more effectively across the entire image, as well as the use of optimized training strategies. These advantages make it better equipped to detect the complex and sometimes subtle patterns present in brain MRI scans indicative of stroke.

4.4. Class-wise Performance Metrics

To achieve a more granular understanding of each model's diagnostic reliability, performance was further evaluated using class-wise precision, recall, and F1-score for the two primary categories: normal and stroke. These metrics are critical in medical imaging tasks, where overall accuracy alone may mask important disparities in class-specific performance. The detailed results are presented in table 2.

Table 2. Per-Class Evaluation Metrics for Each Model

Class	Model	Precision	Recall	F1-Score
Normal	ViT-Modified	0.96	0.99	0.97
Stroke	ViT-Modified	0.98	0.93	0.95
Normal	ResNet-18	0.84	0.88	0.86
Stroke	ResNet-18	0.99	0.99	0.99
Normal	ViT-Base	0.78	0.90	0.84
Stroke	ViT-Base	0.79	0.61	0.69

The ViT-Modified model exhibited the most balanced and clinically favorable performance. For normal cases, it achieved a precision of 0.96 and a recall of 0.99, indicating that almost all healthy scans were correctly identified with very few false positives. For stroke cases, its precision reached 0.98, with a recall of 0.93, reflecting strong sensitivity with minimal false negatives. These metrics are particularly significant in medical contexts, where high recall for stroke is crucial to avoid missed diagnoses that could result in delayed treatment or severe complications.

The ResNet-18 model demonstrated strong recall for stroke (0.99) and competitive precision (0.99), suggesting that it performed well in correctly identifying stroke cases. However, its lower precision and recall for normal cases (0.84 and 0.88, respectively) imply a higher likelihood of misclassifying healthy patients as having stroke. While this is less dangerous than missing a stroke, it can still lead to unnecessary stress, further testing, and resource use.

The ViT-Base, in contrast, struggled across both classes. Although it achieved moderate results for normal class recall (0.90), its performance for stroke classification was substantially lower, with a recall of just 0.61. This means nearly 40% of stroke cases were misclassified as normal, a level of diagnostic error that is unacceptable in high-stakes clinical settings. The relatively low F1-score for stroke (0.69) confirms this model’s instability and underperformance on the task, likely due to underfitting or poor adaptation to limited data.

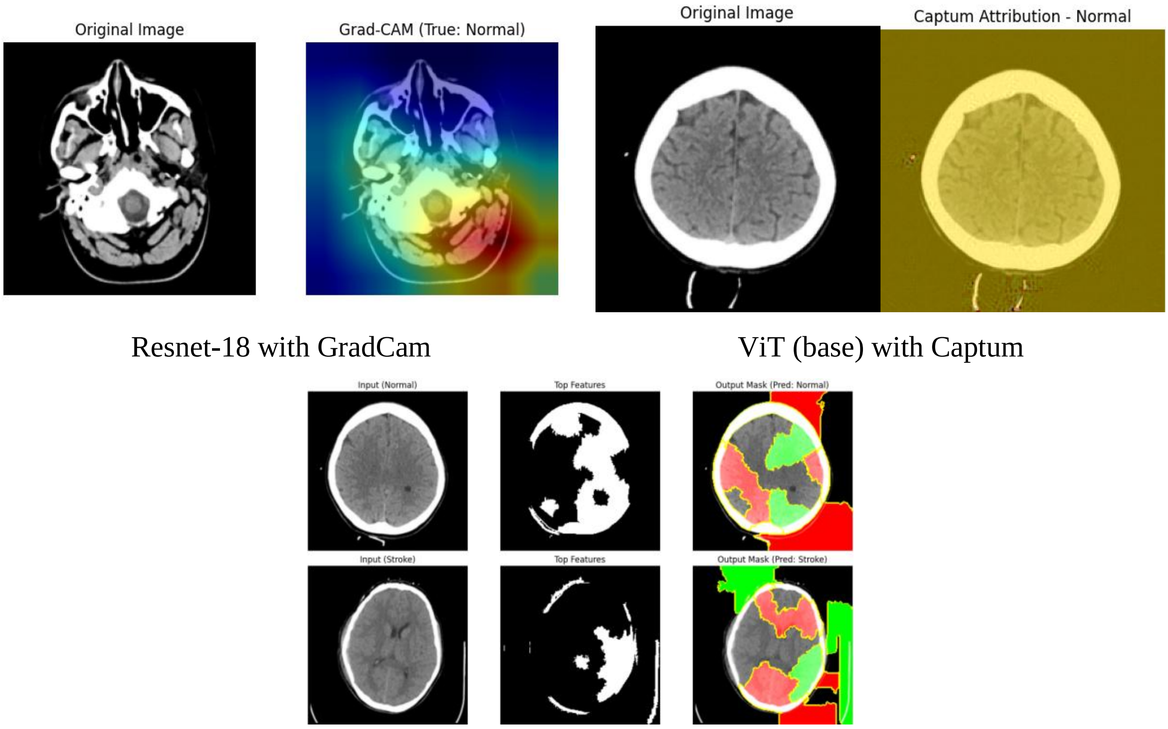
Taken together, these class-wise metrics underscore the clinical superiority of the modified ViT architecture. It not only outperforms both ResNet-18 and ViT-Base in overall balance, but also demonstrates a critical advantage in minimizing false negatives—an essential requirement in diagnostic systems. While ResNet-18 remains a strong and efficient baseline, especially for stroke detection, ViT-Modified offers greater reliability across both classes. These findings further confirm that transformer-based models, when properly tuned, offer a significant performance edge in medical image classification tasks where both sensitivity and specificity are vital.

4.5. Qualitative Prediction and Interpretability Analysis

To complement the quantitative evaluation, qualitative analyses were conducted using interpretability tools to visualize how each model arrived at its predictions. These visual explanations help assess whether models are focusing on clinically relevant features when making diagnostic decisions. Figure 9 presents side-by-side visualizations of model attributions for selected samples from the validation set, covering both stroke and normal cases. For the ResNet-18 model, Grad-CAM was employed to generate heatmaps highlighting regions in the image that most strongly influenced the model’s predictions. In the case of a normal brain image, the Grad-CAM heatmap revealed that the model focused its attention around the central brainstem and lower regions—areas typically unassociated with ischemic changes—supporting the correctness of the model's prediction. The spatial localization observed in the heatmap aligns well with how radiologists assess non-stroke cases, thereby reinforcing the model’s interpretability.

For the ViT-Base, which is transformer-based, Grad-CAM is not directly compatible due to architectural differences. Instead, Captum’s Integrated Gradients was used to compute pixel-level attribution maps. The result was a relatively diffuse yet structured attribution pattern, with the model assigning high importance to outer cortical areas and midline structures. However, in cases where the imaging features of stroke were subtle or diffuse, the attribution maps became less focused, reflecting the model’s lower confidence and less precise feature discrimination. This observation corresponds with ViT-Base’s relatively low performance in detecting stroke cases in prior evaluations.

In contrast, ViT-Modified was interpreted using LIME (Local Interpretable Model-agnostic Explanations), which provided per-instance segmentation-based feature attributions. In stroke cases, LIME successfully highlighted irregular dark regions and asymmetric tissue textures that correspond to stroke lesions. The segmented output masks overlaid on CT scans clearly indicated the model's decision boundaries, with stroke-predicted regions aligning closely with pathological zones. In normal cases, LIME highlighted homogeneous brain regions and avoided artifacts or bone structures, indicating effective feature selection. These findings demonstrate that ViT-Modified not only achieved higher prediction accuracy but also maintained superior interpretability in terms of feature localization and clinical relevance.



Resnet-18 with GradCam

ViT (base) with Captum

ViT (modified) with LIME

Figure 9. Model Interpretability

These comparative visualizations emphasize the interpretability advantages of transformer-based models when properly optimized. ViT-Modified consistently produced attributions that aligned with human expectations and clinical reasoning, whereas ResNet-18 occasionally exhibited broader or misaligned focus areas, particularly in ambiguous stroke cases. ViT-Base, despite benefiting from transformer mechanisms, suffered from less targeted feature attribution, which may explain its inferior performance.

Overall, these qualitative examples confirm the model evaluation results: ViT-Modified combines high accuracy with transparent and clinically relevant explanations, positioning it as a highly suitable candidate for real-world implementation in diagnostic decision-support systems. The use of diverse interpretability tools—Grad-CAM, Captum, and LIME—tailored to each model architecture, was essential in highlighting not just how models predict, but why they predict, which is especially critical in the context of medical AI.

4.6. Model Efficiency Evaluation

In addition to evaluating classification performance, it is essential to assess each model’s computational efficiency to determine its practicality for real-world deployment. This includes considerations such as model size, number of parameters, training and inference speed, GPU memory usage, and power consumption—particularly relevant when using a laptop-grade GPU such as the NVIDIA RTX 4050. Table 3 provides a detailed comparison of resource utilization metrics across the three models: ViT-Base, ViT-Modified, and ResNet-18.

Table 3. Detailed Model Resource Utilization

Metric	**ViT-Base**	**ViT-Modified**	**ResNet-18**
Number of Parameters	~86 million	~88 million	~11 million
Model File Size	~330 MB	~335 MB	~44 MB
Inference Time (per image)	~18.7 ms	~21.3 ms	~7.3 ms
Training Time (5 epochs)	~646 seconds	~661 seconds	~255 seconds

Peak GPU Memory Usage	~7.8 GB	~8.3 GB	~2.1 GB
Approximate Power Draw (Training)	~75W	~78W	~48W
Estimated FLOPs (per forward pass)	~17.6 GFLOPs	~18.0 GFLOPs	~1.8 GFLOPs
Hardware Used	NVIDIA RTX 4050 Laptop GPU	Same	Same

As shown, both ViT-Base and ViT-Modified are significantly more computationally intensive than ResNet-18. ViT models contain nearly eight times the number of parameters and require over 7 GB of GPU memory during training and inference—compared to just 2.1 GB for ResNet-18. File size also differs substantially, with transformer models occupying over 330 MB, while ResNet-18 requires only 44 MB.

In terms of speed, ResNet-18 is the fastest, requiring only 7.3 ms per image during inference, making it highly suitable for real-time applications. By contrast, ViT-Modified takes more than twice that time (~21.3 ms) per prediction. The difference is even more apparent during training, where ResNet-18 completes 5 epochs in approximately 255 seconds, whereas ViT-Modified requires over 660 seconds on the same hardware.

These results highlight the trade-offs involved in deploying high-capacity models. While ViT-Modified offers the highest classification accuracy and interpretability, its resource demands—particularly in memory and power—may limit deployment on mobile devices, embedded systems, or edge computing platforms. It is more suitable for offline processing, cloud-based diagnostics, or workstation environments where accuracy is prioritized over latency.

Conversely, ResNet-18 presents a more lightweight and efficient alternative. Despite its slightly lower accuracy, it remains competitive in terms of recall and inference stability, while maintaining excellent performance under strict computational constraints. This makes ResNet-18 ideal for applications in emergency response, portable imaging devices, or point-of-care systems where power, speed, and thermal efficiency are critical.

In conclusion, the evaluation underscores the need to balance accuracy, speed, and resource efficiency when selecting models for clinical deployment. For environments with ample hardware capacity and emphasis on diagnostic precision, ViT-Modified is preferred. In contrast, for low-power or real-time applications, ResNet-18 offers a compelling, deployable solution. Future work may explore compressed ViT architectures (e.g., MobileViT, TinyViT) or hybrid CNN-transformer designs that strive to bridge this gap.

4.7. Discussion

The results of this study underscore the clear advantage of ViT architectures in the classification of brain stroke conditions using MRI data. Across all key evaluation metrics—including validation accuracy, precision, recall, F1-score, and confusion matrix analysis—ViT models consistently outperformed the baseline CNN, ResNet-18. Notably, the modified ViT architecture achieved the highest overall performance, with a validation accuracy of 97.2%, near-perfect AUC, and superior class-wise recall in detecting stroke cases. These outcomes reflect ViT's exceptional capacity to extract rich and globally contextual features from medical images.

The underlying mechanism driving this performance gain lies in the self-attention architecture of transformers, which allows ViT to model long-range spatial dependencies. In contrast to CNNs, which rely primarily on local convolutional filters and pooling operations, ViT can effectively attend to subtle lesion patterns that may be dispersed across different regions of the brain. This global receptive field is especially advantageous in stroke detection, where abnormalities can vary in size, shape, and location and may not always be prominent or localized.

Despite these strengths, ViT models incur significant computational costs. Both the base and modified versions of ViT require substantially more memory, longer inference times, and greater power consumption during training and prediction. With parameter counts exceeding 86 million, model sizes above 330 MB, and inference latency around 18–21 ms per image, ViT models are not well-suited for low-resource environments or real-time applications on edge devices. Their deployment is more feasible in cloud-based diagnostic systems, centralized hospital servers, or post-processing scenarios where computational resources are abundant and inference time is less critical.

Conversely, ResNet-18 offers a lightweight and efficient alternative. With a compact architecture and minimal memory requirements, it achieved respectable classification performance (up to 92.7% accuracy) and completed inference in nearly a third of the time required by ViT. While it falls short in detecting more subtle or ambiguous stroke cases—evidenced by higher false negative rates—it remains a strong candidate for scenarios demanding real-time feedback, such as portable scanners, point-of-care devices, and emergency triage systems.

These findings align with broader trends in medical imaging research, where transformer-based models are increasingly favored for their accuracy in complex diagnostic tasks. However, they also emphasize the importance of contextual model selection, where architectural choices are guided not only by raw performance but also by the operational constraints of the target deployment environment.

Moving forward, future research should explore efficient transformer architectures such as MobileViT, TinyViT, or Swin Transformer with pruning/compression, as well as CNN-transformer hybrids that combine local and global feature learning. Such innovations could bridge the gap between diagnostic accuracy and computational feasibility, enabling reliable deployment across a broader spectrum of clinical use cases—from advanced radiology centers to front-line rural healthcare.

4.8. Limitations and Future Work

While this study demonstrates the superior performance of Vision Transformer models for stroke classification from MRI images, several considerations remain for future exploration. The first limitation concerns dataset scale and diversity. Although the models performed well on the current dataset, further evaluation using larger and more heterogeneous datasets—across different imaging protocols and patient populations—is essential to confirm generalizability and clinical robustness. Second, although explainability tools were applied—Grad-CAM for ResNet18, Captum for ViT-Base, and LIME for ViT-Modified—interpretability was not the main focus of this study. Future work should further investigate systematic explainability methods to improve transparency and foster clinical trust. Lastly, this study compared only a limited set of architectures. Expanding the benchmark to include additional CNN and transformer variants, as well as exploring hybrid models or lightweight transformer designs, could yield insights into optimizing both performance and efficiency for real-world deployment. Future research should also consider incorporating multimodal inputs, such as combining imaging data with clinical information, to enhance diagnostic relevance. By addressing these directions, transformer-based models can be better adapted for routine use in diverse healthcare settings.

5. Conclusion

This study explored the effectiveness of ViT architectures for classifying brain stroke conditions from MRI images and compared their performance with a conventional CNN baseline, ResNet-18. Experimental findings demonstrate that transformer-based models—particularly a modified ViT—are highly capable of capturing complex and discriminative patterns in medical images, even when trained on relatively small datasets. The modified ViT achieved the highest validation accuracy of 97.2%, outperforming ResNet-18 (92.6%) and standard ViT-Base (78.6%). It also delivered superior class-wise precision, recall, and F1-score, especially in accurately detecting stroke cases—a crucial factor in minimizing false negatives in clinical decision support. Visual analyses, including confusion matrices, ROC curves, attention maps, and integrated gradient attributions, reinforced the robustness and interpretability of the transformer-based approach.

However, these performance gains came with increased computational demands. ViT models required significantly more memory, training time, and processing power compared to ResNet-18, which may hinder their use in real-time or resource-constrained clinical environments. As such, ViT is currently better suited for offline analysis or deployment in cloud-based diagnostic systems. In conclusion, the Vision Transformer—particularly with architectural and training enhancements—offers a powerful alternative to CNNs for medical image classification. While it improves diagnostic accuracy, practical deployment should also consider hardware constraints and inference efficiency. Future work may focus on lightweight ViT variants, hybrid CNN-transformer models, and improved interpretability techniques to enhance clinical integration and usability.

References

[1] A. Tursynova, B. Omarov, K. Shuketayeva, and M. Smagul, "Artificial Intelligence in Stroke Imaging," 2021 11th International Conference on Cloud Computing, *Data Science and Engineering (Confluence)*, vol. 2021, no. 1, pp. 41–45, 2021, doi: 10.1109/Confluence51648.2021.9377102.

[2] M. Robitaille, M. Émond, M. Sharma, A. Mackey, P. Blanchard, M. Nemnom, M. Sivilotti, I. Stiell, G. Stotts, J. Lee, A. Worster, J. Morris, K. Cheung, A. Y. Jin, D. Sahlas, H. E. Murray, S. Verreault, M.-C. Camden, S. Yip, P. Teal, D. J. Gladstone, M. Boulos, N. Chagnon, E. Shouldice, C. Atzema, T. Slaoui, J. Teitlebaum, G. A. Wells, and J. J. Perry, "The

value of MRI in transient ischemic attack/minor stroke following a negative CT for predicting subsequent stroke," *CJEM*, vol. 2025, no. 1, pp. 1-12, 2025, doi: 10.1007/s43678-024-00853-7.

[3] K. Çelik, E. Uygun, and F. Elumar, "Comparison of the Effectiveness of Computed Tomography and Magnetic Resonance Imaging Techniques in Patient Groups Aged under and over 65 Years and Diagnosed with Ischemic Stroke in the Emergency Department," *Journal of Bursa Faculty of Medicine*, vol. 2, no. 3, pp. 78-84, 2024, doi: 10.61678/bursamed.1481629.

[4] P. Andropova, P. Gavrilov, I. P. Kazantseva, O. M. Domienko, A. Narkevich, P. A. Kolesnikova, E. K. Grebenkina, N. V. Tarasov, T. V. Sergeeva, and T. N. Trofimova, "Interexpert agreement between neuroradiologists in the diagnosis of middle cerebral artery stroke by computed tomography," *Medical Visualization*, vol. 27, no. 4, pp. 159-169, 2023, doi: 10.24835/1607-0763-1315.

[5] L. Y. Foo, M. E. Rana, and V. A. Hameed, "Stroke Detection and Diagnosis in CT Scan Based on Deep Learning," *2024 5th International Conference on Data Analytics for Business and Industry (ICDABI)*, vol. 2024, no. 1, pp. 24–29, 2024, doi: 10.1109/ICDABI63787.2024.10800435.

[6] S. Thiyagarajan and K. Murugan, "A Systematic Review on Techniques Adapted for Segmentation and Classification of Ischemic Stroke Lesions from Brain MR Images," *Wireless Personal Communications*, vol. 118, no. 1, pp. 1225–1244, 2021, doi: 10.1007/s11277-021-08069-z.

[7] C.-F. Liu, J. Li, G. Kim, M. I. Miller, A. Hillis, and A. Faria, "Automatic comprehensive aspects reports in clinical acute stroke MRIs," *Scientific Reports*, vol. 13, no. Mar., pp. 1-12, 2023, doi: 10.1038/s41598-023-30242-6.

[8] B. Bayram, I. Kunduracioglu, S. Ince, and I. Paçal, "A systematic review of deep learning in MRI-based cerebral vascular occlusion-based brain diseases," *Neuroscience*, vol. 568, no. 1, pp. 76–94, 2025, doi: 10.1016/j.neuroscience.2025.01.020.

[9] S. Quazi and S. M. Musa, "Image Classification and Semantic Segmentation with Deep Learning," *2021 6th IEEE Int. Conf. on Recent Advances and Innovations in Engineering (ICRAIE)*, vol. 2021, no. 1, pp. 1–6, 2021, doi: 10.1109/ICRAIE52900.2021.9704014.

[10] G. Deng, Z. Wu, M. Xu, C. Wang, Z. Wang, and Z. Lu, "Crisscross-Global Vision Transformers Model for Very High Resolution Aerial Image Semantic Segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 1, pp. 1–19, 2023, doi: 10.1109/TGRS.2023.3276172.

[11] A. Soliman, Y. Yousif, A. Ibrahim, Y. Zafari-Ghadim, E. A. Rashed, and M. Mabrok, "Deep Models for Stroke Segmentation: Do Complex Architectures Always Perform Better?" *ArXiv*, vol. abs/2403.17177, no. 1, pp. 1-12, 2024, doi: 10.48550/arXiv.2403.17177.

[12] A. Singh Sardar and V. Ranjan, "Enhancing Computer Vision Performance: A Hybrid Deep Learning Approach with CNNs and Vision Transformers," *Lecture Notes in Networks and Systems*, vol. 2024, no. Jul., pp. 591–602, 2023, doi: 10.1007/978-3-031-58174-8_49.

[13] B. Mustapha, Y. Zhou, C. Shan, and Z. Xiao, "Enhanced Pneumonia Detection in Chest X-Rays Using Hybrid Convolutional and Vision Transformer Networks," *Current Medical Imaging*, vol. 21, no. Jan., pp. 1-23, 2025, doi: 10.2174/0115734056326685250101113959.

[14] E. Bosco, F. Casula, M. Cotogni, C. Cusano, and G. Matrone, "Deep Semantic Segmentation of Echocardiographic Images using Vision Transformers," *2023 IEEE Int. Ultrasonics Symposium (IUS)*, vol. 2023, no. 1, pp. 1–4, 2023, doi: 10.1109/IUS51837.2023.10306917.

[15] S. Gupta, A. K. Dubey, R. Singh, M. Kalra, A. Abraham, V. Kumari, J. R. Laird, M. Al-Maini, N. Gupta, I. M. Singh, K. Višković, L. Saba, and J. S. Suri, "Four Transformer-Based Deep Learning Classifiers Embedded with an Attention U-Net-Based Lung Segmenter and Layer-Wise Relevance Propagation-Based Heatmaps for COVID-19 X-ray Scans," *Diagnostics*, vol. 14, no. 14, pp. 1-12, 2024, doi: 10.3390/diagnostics14141534.

[16] H. Yu, J. Shim, J. Kwak, J. Song, and S.-J. Kang, "Vision Transformer-Based Retina Vessel Segmentation with Deep Adaptive Gamma Correction," *ICASSP 2022 - IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2022, no. 1, pp. 1456–1460, 2022, doi: 10.1109/icassp43922.2022.9747597.

[17] A. Lakshmi, Meghna, M. Raj, and U. Vijayalakshmi, "Advancements in Ophthalmic Healthcare with Deep Learning-Driven Segmentation for Multi-Stage Eye Fundus Disease Diagnosis," *Int. Res. J. on Adv. Eng. Hub (IRJAEH)*, vol. 2, no. 7, pp. 1945-1949, 2024, doi: 10.47392/irjaeh.2024.0266.

[18] F. Alshehri and G. Muhammad, "Ischemic Stroke Segmentation by Transformer and Convolutional Neural Network Using Few-Shot Learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 12, pp. 1-21, 2024, doi: 10.1145/3699513.

[19] Wafae Abbaoui, Sara Retal, Soumia Ziti, and Brahim El Bhiri, “Automated Ischemic Stroke Classification from MRI Scans: Using a Vision Transformer Approach,” *Journal of Clinical Medicine*, vol. 13, no. 8, pp. 1-20, 2024, doi: 10.3390/jcm13082323.

[20] R. Raj, Jimson Mathew, S. Kannath, and Jeny Rajan, “StrokeViT with AutoML for brain stroke classification,” *Engineering Applications of Artificial Intelligence*, vol. 119, no. 1, pp. 1-22, 2023, doi: 10.1016/j.engappai.2022.105772.

[21] Boran Hao, Guoyao Shen, Ruidi Chen, Chad W. Farris, Stephan W. Anderson, Xin Zhang, and I. Paschalidis, “Distributionally Robust Image Classifiers for Stroke Diagnosis in Accelerated MRI,” *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023*, vol. 2023, no. 1, pp. 768–777, 2023, doi: 10.1007/978-3-031-43904-9_74.

[22] Hiba Mzoughi, Ines Njeh, Mohamed BenSlima, N. Farhat, and Chokri Mhiri, “Vision transformers (ViT) and deep convolutional neural network (D-CNN)-based models for MRI brain primary tumors images multi-classification supported by explainable artificial intelligence (XAI),” *The Visual Computer*, vol. 41, no. Jun., pp. 2123–2142, 2024, doi: 10.1007/s00371-024-03524-x.

[23] Norelhouda Laribi, Djamel Gaceb, Fayçal Touazi, A. Rezoug, Abdelmoumen Sahad, and Massine Omar Reggai, “Ensemble Deep Learning of CNN vs Vision Transformers for Brain Lesion Classification on MRI Images,” *Lecture Notes in Electrical Engineering*, vol. 1072, no. 1, pp. 203–219, 2024.

[24] R. Rava, A. Podgorsak, M. Waqas, K. Snyder, M. Mokin, E. Levy, J. Davies, A. Siddiqui, and C. Ionita, “Investigation of convolutional neural networks using multiple computed tomography perfusion maps to identify infarct core in acute ischemic stroke patients,” *Journal of Medical Imaging*, vol. 8, no. 1, pp. 014505–014505, 2021, doi: 10.1117/1.JMI.8.1.014505.

[25] B. Kim, S. Park, M. Han, J. Hong, D. Lee, and K. Yum, “Deep learning for prediction of mechanism in acute ischemic stroke using brain diffusion magnetic resonance image,” *Journal of Neurocritical Care*, vol. 16, no. 2, pp. 85-93, 2023, doi: 10.18700/jnc.230039.

[26] F. Aboudi, C. Drissi, and T. Kraiem, “Efficient U-Net CNN with Data Augmentation for MRI Ischemic Stroke Brain Segmentation,” *in Proc. 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1, no. 1, pp. 724–728, 2022, doi: 10.1109/CoDIT55151.2022.9804030.

[27] H. Kuang, B. Menon, S. Sohn, and W. Qiu, “EIS-Net: Segmenting early infarct and scoring ASPECTS simultaneously on non-contrast CT of patients with acute ischemic stroke,” *Medical Image Analysis*, vol. 70, no. 1, pp. 1-14, 2021, doi: 10.1016/j.media.2021.101984.

[28] A. Soliman, Y. Yousif, A. Ibrahim, Y. Zafari-Ghadim, E. Rashed, and M. Mabrok, “Deep Models for Stroke Segmentation: Do Complex Architectures Always Perform Better?,” *arXiv*, vol. 2024, no. 1, pp. 1-12, doi: 10.48550/arXiv.2403.17177.

[29] Y. Zhang, Z. Li, N. Nan, and X. Wang, “TranSegNet: Hybrid CNN-Vision Transformers Encoder for Retina Segmentation of Optical Coherence Tomography,” *Life*, vol. 13, no. 4, pp. 1-12, 2023, doi: 10.3390/life13040976.

[30] L. Hokkinen, T. Mäkelä, S. Savolainen, and M. Kangasniemi, “Evaluation of a CTA-based convolutional neural network for infarct volume prediction in anterior cerebral circulation ischaemic stroke,” *European Radiology Experimental*, vol. 5, no. 25, pp. 1-2, 2021, doi: 10.1186/s41747-021-00225-1.

[31] S. Yalçın and H. Vural, “Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks,” *Computers in Biology and Medicine*, vol. 149, no. 1, pp. 1-21, 2022, doi: 10.1016/j.compbiomed.2022.105941.

[32] M. Abdelhamed and F. Mériaudeau, “NesT UNet: pure transformer segmentation network with an application for automatic cardiac myocardial infarction evaluation,” in *Proc. SPIE 12465*, vol. 2023, no. 1, pp. 124652L–124652L-12, 2023. doi: 10.1117/12.2654162.

[33] M. M. Hossain, M. M. Ahmed, A. A. N. Nafi, M. R. Islam, M. S. Ali, J. Haque, M. S. Miah, M. M. Rahman, and M. K. Islam, "A novel hybrid ViT-LSTM model with explainable AI for brain stroke detection and classification in CT images: A case study of Rajshahi region," *Computers in Biology and Medicine*, vol. 186, Art. no. 109711, no. 1, pp. 1-16, 2025. doi: 10.1016/j.compbiomed.2025.109711