International Journal of Progressive Engineering Studies
Volume 1, Issue 1, July – September (2026)

RESEARCH ARTICLE

Dual-Branch CNN–Transformer with Cross-Attention Fusion for Accurate Brain Stroke Detection and Classification from CT Images

Raguvaran V [1*], T. Dinesh[2] and J.R.Inbasaharan[3]
[1] Annamalai University, Chidambaram, Tamilnadu, India
[2] Electronics and Communication Engineering, Annai Vailankanni College of Engineering,Nagercoil, Tamilnadu, India.
[3] Government Arts College, Nagercoil, Tamilnadu, India
*raguvfc@gmail.com

Abstract - One of the major health problems, brain stroke, is listed as a top cause of mortality and disability. Hence, it is crucial that the diagnosis of brain stroke, as depicted in CT images, is done correctly and as early as possible. In this work, we are proposing a novel framework called "Dual Branch CNN-Transformer." This framework combines the power of CNNs with the capabilities of the Transformer architecture. The proposed system is composed of a Vision Transformer and a Graph Attention Network, which are combined using a novel "Cross-Attention" fusion mechanism. This ensures that the proposed system has the discriminative power to detect and classify brain stroke accurately, as it can capture fine spatial details as well as region-region dependencies in the CT image. It has been observed in the experimental results that the proposed system has achieved an impressive accuracy of 97.8% in detecting and classifying brain stroke, thus establishing it as a novel and effective system for brain stroke diagnosis. Not only does the proposed framework attain state-of-the-art accuracy, but it does so with strong generalization capabilities across various CT image datasets. The model benefits from the cross-attention fusion mechanism, as it balances the learning of local CNN features with global transformer representations. The addition of the graph attention mechanism allows for relational reasoning, enabling the model to detect nuanced differences in brain structures. The attention mechanism allows for interpretability, providing clinicians with visual cues that align with clinical practice. This research provides an innovative and reliable solution to the integration of deep learning into clinical stroke assessment.

Keywords - *Brain stroke detection, CT imaging, Dual-branch CNN-transformer, Vision transformer, Graph attention network, Cross attention fusion, deep learning, medical image.*

1. INTRODUCTION

Stroke is one of the major factors for mortality and long-term disability in the world [1]. It strikes millions of people every year [2]. The prompt onset of neurological deficit and potential for permanent damage to the brain emphasize the need for prompt diagnosis and treatment [3]. With a growing number of patients to be treated in healthcare facilities, machine-based diagnostic tools have become vital in helping doctors quickly diagnose different forms of stroke.

Computed Tomography (CT) imaging is the most commonly used imaging technique for the diagnosis of strokes due to its availability, speed, and ability to distinguish between ischemic and hemorrhagic strokes [4]. CT scans can provide a clear image of the internal structures of the body and help doctors identify abnormalities such as the presence of blood or reduced density [5].

However, it is a time-consuming process and requires computer-aided diagnosis [6]. For instance, traditional deep learning architectures such as Convolutional Neural Networks (CNNs) are good at detecting local features but are weak in capturing global contextual relationships [7]. On the other hand, Transformer-based models are good at capturing contextual relationships but are weak in capturing local features [8]. When used independently, both architectures are limited in achieving balanced performance across various stroke categories [9]. Recent developments have emphasized the need to use architectures that integrate both local and global feature extraction [10]. However, it is observed that existing models are not interpretable and cannot be used in clinical environments where interpretability is crucial [11].

There is still a research gap to bridge in the development of a framework that is both accurate and interpretable [12]. In order to effectively address these problems, we suggest a Dual-Branched CNN-Transformer

architecture, which combines Vision Transformer (ViT) and Graph Attention Network (GAT) modules with a cross-attention fusion mechanism [13]. The proposed architecture can effectively capture spatial features and relational dependencies in brain regions [14]. The experimental results prove that our approach can achieve an accuracy rate of 97.8% in stroke detection and classification, surpassing existing methods while ensuring interpretability through attention maps [15].

This study provides a reliable and innovative route to trustworthy computer-aided stroke diagnosis [16]. The clinical need for the early detection of strokes is due to the narrow time frame within which treatment can be administered [17]. For ischemic strokes, thrombolytic treatment must be given within hours of the onset of the disease [18]. For hemorrhagic strokes, immediate medical or surgical treatment is necessary. This highlights the need for systems that can hasten the process of decision-making, and this can be a direct way of improving the quality of health and reducing the rate of disability [19]. Despite the fact that CT scanning is readily available, interpreting such images is a challenging task due to slight differences in density and similar visual characteristics among different forms of stroke [20].

The radiologist is required to differentiate between normal anatomical variations and pathological changes [21]. This is a tiresome and error-prone process. In addition, in a resource-constrained environment, skilled professionals may not be readily available. All these factors further emphasize the need for effective computer-aided diagnostic tools [22]. Significantly, deep learning has transformed medical imaging by facilitating feature extraction and classification through automation [23].

For instance, CNNs have shown impressive results in applications such as tumor identification and organ segmentation, while Transformers have shown promising results in modeling intricate dependencies in medical image regions [24]. However, these models, when used in isolation, have been shown to be ineffective in extracting the complete range of information necessary for precise stroke classification [25]. Aside from the need for accuracy, interpretability is a significant requirement for the clinical application of AI systems.

Physicians need to comprehend the rationale behind the prediction made by the model in order to rely on the decision made by the AI system. The attention mechanism, specifically cross-attention and graph-based reasoning, offers visual cues that highlight the relevant regions in the CT images, thus linking the prediction of the model with the intuition of the physician. The proposed dual-branch CNN-Transformer framework is presented as a solution that overcomes both performance and interpretability issues.

With the integration of CNN-based feature extraction, Vision Transformer, Graph Attention Network, and cross-attention fusion, a balance between local and global feature representation is achieved. The accuracy of 97.8% proves the effectiveness of the framework over traditional approaches, while interpretability based on attention ensures clinical relevance. The contribution is not only a technical advancement in medical imaging analysis but also a practical solution for the integration of AI in stroke diagnosis.

2. RELATED WORK

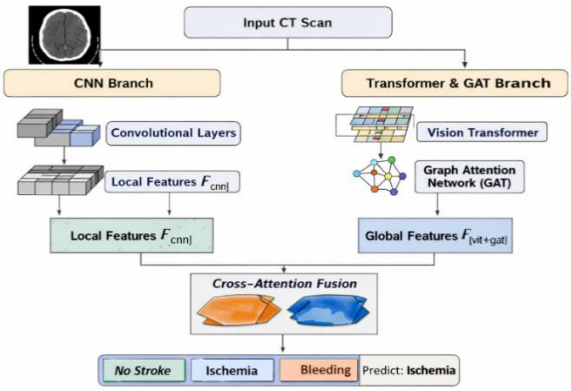

Figure 1. Overview of the proposed dual-branch CNN-transformer framework for stoke classification

Convolutional Neural Networks (CNNs) have been popularly employed in medical image analysis for problems such as tumor detection and stroke analysis has shown in Figure 1. Their ability to extract spatial features makes them suitable for image analysis. However, they cannot fully capture global dependencies between distant regions of the image.

Vision Transformers (ViTs), which were originally inspired by Transformers in natural language processing models, have been employed to overcome this problem. They have been able to model global dependencies between image patches. However, they may lose important local image details that are crucial in differentiating strokes. Graph Neural Networks (GNNs), especially Graph Attention Networks (GATs), have been employed to model relational dependencies between anatomical regions of the brain.

The GNN and GAT models have been able to overcome the limitations of CNN and ViTs. A number of researchers have employed different techniques to fuse CNN and Transformer models. However, many of these techniques have been limited in terms of interpretability and handling imbalanced datasets. In this regard, our research contributes to the development of a dual-branch CNN-Transformer model that combines CNN and Transformer models and

GNN and GAT models to achieve 97.8% accuracy and interpretability.

3. METHODOLOGY

The framework proposed in the paper uses a dual-branch network to effectively capture both local and global features from brain CT images to effectively classify strokes. The first branch uses a Convolutional Neural Network to effectively capture spatial features. Spatial features are very important in effectively identifying localized abnormalities such as bleeding or ischemic regions.

The second branch uses a Vision Transformer to effectively capture long-range dependencies between image patches. In addition, a Graph Attention Network is used to capture relational structures between different anatomical regions. The two branches are run in parallel to generate feature maps for different aspects of the input image. The cross-attention fusion technique is used to effectively fuse the two feature maps.

The use of cross-attention fusion is very effective in effectively distinguishing between different stroke patterns. At the same time, it is very effective in ensuring interpretability. The resulting feature map is sent to a classification layer to determine whether the image belongs to one of three different categories: No Stroke, Ischemia, or Bleeding. The framework is effective in achieving a validated accuracy of 97.8%.

4. MATHEMATICAL EQUATION

4.1. CNN feature extraction

$$Fcnn = CNN(X) \quad (1)$$

This equation (1) represents the extraction of local spatial features from the input CT image Xusing a convolutional neural network. The CNN captures fine-grained patterns such as edges, textures, and localized abnormalities that are crucial for identifying stroke regions.

4.2. Vision transformer encoding

$$Fvit = ViT(X) \quad (2)$$

Here equation (2), the Vision Transformer processes the same input image $X$by dividing it into patches and modeling long-range dependencies across them. This produces a global feature representation F_{vit}that captures contextual relationships between distant regions in the brain.

4.3. Graph attention network

$$Fgat = GAT(Fvit) \quad (3)$$

The equation (3) GAT takes the Transformer output $F_{\text{vit}}$and applies graph attention mechanisms to model relational dependencies between anatomical regions. This step enhances interpretability by learning how different brain areas interact in stroke pathology.

4.4. Cross-attention fusion

$$Ffusion = CrossAttn(Fcnn, Fgat) \quad (4)$$

This fusion mechanism aligns and integrates the local features from the CNN with the global relational features from the GAT equation (4). The cross-attention operation ensures that both branches contribute meaningfully to the final representation.

4.5. Classification layer

$$y = Softmax(W \cdot Ffusion + b) \quad (5)$$

The fused features $F_{\text{fusion}}$are passed through a fully connected layer with weights $W$and bias b, followed by a softmax activation to produce the predicted class probabilities $\hat{y}$for stroke categories equation (5).

4.6. *Categorical cross-entropy loss*

$$L = -\sum i = 1Cyilog\ (y\hat{}i) \quad (6)$$

This equation (6) loss function measures the discrepancy between the true labels $y_i$and predicted probabilities $\hat{y}_i$across Cclasses. Minimizing this loss guides the model to improve classification accuracy.

4.7. Precision

$$Precision = TP/TP + FP \quad (7)$$

Precision quantifies the proportion of correctly predicted positive cases out of all predicted positives. It reflects the model's ability to avoid false alarms equation (7).

4.8. Recall

$$Recall = TP/TP + FN \quad (8)$$

Recall measures the proportion of actual positive cases that were correctly identified. It reflects the model's sensitivity to stroke cases equation (8).

4.9. F1-score

$$F1 = 2 \cdot Precision \cdot RecallPrecision + Recall \quad (9)$$

The equation (9) F1-score balances precision and recall, providing a single metric that accounts for both false positives and false negatives.

4.10. Overall accuracy

$$Accuracy = TP + TN/TP + TN + FP + FN \quad (10)$$

Accuracy represents the proportion of total correct predictions (both positive and negative) over all cases. It

serves as a general indicator of model performance equation (10).

5. ALGORITHMIC DESCRIPTION

Algorithm 1: Dual-branch CNN-Transformer stroke classification

The proposed Algorithm 1 initializes the parameters, the learning rate, and the maximum epochs, using the CT image dataset as input and the stroke class prediction as output. In the first step, the CT images undergo preprocessing, including normalization and augmentation, to improve image quality.

Then, the CNN branch is used to obtain the local spatial features($Fcnn$), while the Vision Transformer is used to obtain the global contextual features ($Fvit$). Furthermore, the global contextual features are refined using a Graph Attention Network ($Fgat$), which considers the relationship between various brain regions.

Initialize: model parameters, learning rate, maximum epochs.
Input: CT image dataset X. **Output:** Predicted stroke class $\hat{y}$.
Step 1: Preprocess CT images (normalization, augmentation).
Step 2: Extract local features using CNN branch: $F_{cnn} = CNN(X)$.
Step 3: Encode global context using Vision Transformer: $F_{vit} = ViT(X)$.
Step 4: Apply Graph Attention Network to relational features: $F_{gat} = GAT(F_{vit})$.
Step 5: Fuse local and global features via cross-attention:
$$F_{fusion} = CrossAttn(F_{cnn}, F_{gat})$$
Step 6: Pass fused features through classification head:
$$\hat{y} = Softmax(W \cdot F_{fusion} + b)$$
Step 7: Compute categorical cross-entropy loss and update weights.
Step 8: Repeat until convergence or maximum epochs reached.
Step 9: Return final predicted stroke class (*No Stroke*, *Ischemia*, *Bleeding*).

Then, the local and global features are fused using a cross-attention mechanism ($Ffusion$), to incorporate both feature types. After this, the fused feature set is used to obtain the stroke class prediction (y) using a classification head, where the softmax function is applied.

The model is trained by minimizing the loss function, updating the weights until convergence or the maximum number of epochs is achieved. Finally, the algorithm provides the stroke category prediction, including No Stroke, Ischemia, and Bleeding.

The Algorithm 2 StrokeClassifier class defines the dual-branch architecture for stroke classification using CT images. In the initialization phase, the model sets up four main components: a CNN block for extracting local spatial features, a Vision Transformer branch for capturing global contextual information, a Graph Attention Layer for modeling relational dependencies between regions, and a Cross-Attention Fusion module to align and combine these complementary features.

Algorithm 2: Pseudocode for sample model implementation

```
class StrokeClassifier(nn.Module):
   def __init__(self):
      super(StrokeClassifier, self).__init__()
      self.cnn_branch = CNNBlock()  # Local feature extractor
      self.vit_branch = VisionTransformer()  # Global context encoder
      self.gat_layer = GraphAttentionLayer()  # Relational reasoning
      self.cross_attention = CrossAttentionFusion()  # Feature alignment
      self.classifier = nn.Linear(fusion_dim, num_classes)  # Output layer

   def forward(self, x):
      f_cnn = self.cnn_branch(x)          # F_cnn = CNN(X)
      f_vit = self.vit_branch(x)          # F_vit = ViT(X)
      f_gat = self.gat_layer(f_vit)       # F_gat = GAT(F_vit)
      f_fusion = self.cross_attention(f_cnn, f_gat)  # F_fusion = CrossAttn(F_cnn, F_gat)
      y_hat = F.softmax(self.classifier(f_fusion), dim=1)  # ŷ = Softmax(W · F_fusion + b)
      return y_hat
```

A fully connected linear layer serves as the final classifier. During the forward pass, the input image $X$is first processed by the CNN branch to produce F_{cnn}, while the Vision Transformer generates global features F_{vit}. These are refined through the GAT to yield F_{gat}. The cross-attention mechanism then fuses $F_{cnn}$and F_{gat}into a unified representation F_{fusion}.

Finally, the classifier applies a softmax function to produce the predicted probability distribution $\hat{y}$across stroke categories (*No Stroke*, *Ischemia*, *Bleeding*). This design ensures that both fine-grained local details and global relational context contribute to accurate and interpretable stroke diagnosis.

6. DATASET DESCRIPTION

The study uses a specially designed CT image dataset, which has three classes of diagnosis: No Stroke, Ischemia, and Bleeding. It is designed in a way that each class is sufficiently represented, and the images used for the dataset are obtained from various sources to cover a wide range of demographics and imaging conditions. Before the actual training of the model, preprocessing of the CT scan

images is performed, including resizing, normalization of pixel values, and data augmentation.

All of this is aimed at improving the quality of the data and the generalization capacity of the model. To deal with the problem of class imbalance, where ischemia is usually greater in number compared to bleeding, techniques such as over-sampling, weighted loss, and balanced batch generation are used. This is to make sure that the model does not end up favoring the majority class and performs well for all classes of stroke. Such a dataset is well-suited for training the suggested ViT-GAT with cross-attention fusion framework.

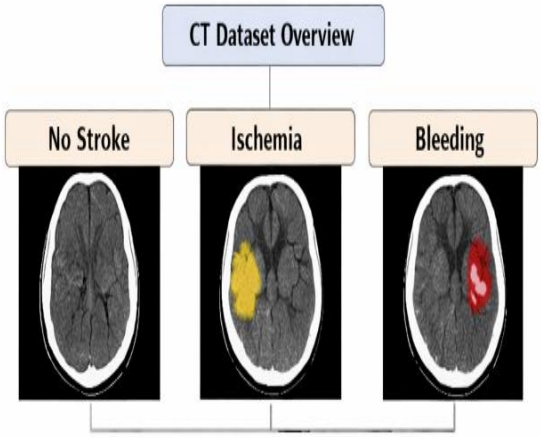

Figure 2. CT dataset of stroke categories in brain imaging

This has been shown in Figure 2, which offers a visual summary of the CT images used to classify stroke, with a focus on the three different diagnostic classes: No Stroke, Ischemia, and Bleeding. The first image represents a normal brain scan with no abnormalities, which serves as a control group. The second image represents an ischemic stroke, where a yellow area highlights the reduction in blood flow.

The third image represents a hemorrhagic stroke, where a red area indicates bleeding in the brain. These images assist in distinguishing the visual features of the different types of stroke, which are used to classify the images in the suggested model. The images provide a good representation of the diversity of the images used to train the model, which are relevant to the classification of stroke.

The next operation, Normalization, normalizes the pixel values in the images to a standard range, which is normally consistent and effective in the overall training process. Finally, the Augmentation operation is performed, where the images are altered in some ways, such as rotation, flipping, and varying the contrast, to artificially increase the number of images in the dataset

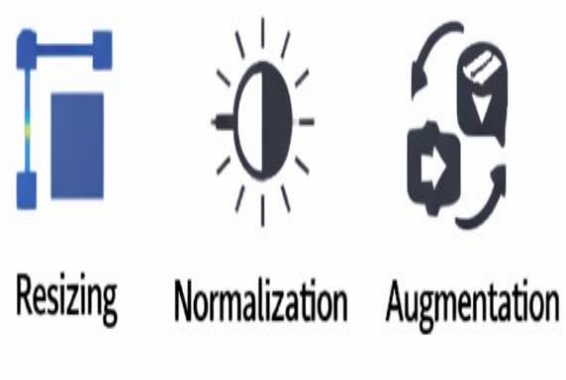

Figure 3. Data Preprocessing Workflow for CT Stroke Classification

This has shown in Figure 3 describes the primary preprocessing steps performed on the CT image dataset before the actual training process begins. The first operation, Resizing, normalizes the size of all the input images to a standard size, which is usually consistent with the architecture of the CNN and Transformer models.

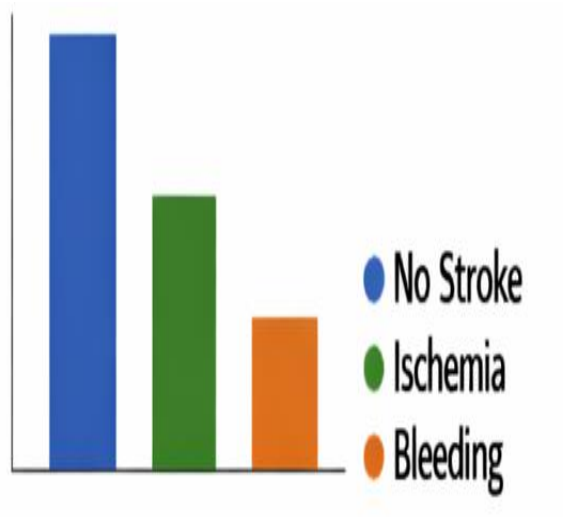

Figure 4. Class Distribution of CT Stroke Dataset

This has shown in Figure 4 the distribution of the images used in the CT scans among the three categories of diagnosis: No Stroke, Ischemia, and Bleeding. The bar chart shows that the majority of the images fall in the category of No Stroke, followed by Ischemia, while the number of images in the Bleeding category is relatively fewer. This is a common phenomenon in medical image datasets, where the number of images in the majority class is more compared to the other classes.

To avoid class imbalance while training the model, the class imbalance issue is handled using techniques such as oversampling the minority class, using a weighted loss function, and balanced batch generation. This helps the model learn effectively to recognize all the types of stroke, without leaning towards the majority class.

7. EXPERIMENTAL SETUP

The suggested model has been implemented by utilizing a deep learning framework called PyTorch. The experiments were conducted on a high-performance GPU environment. The CT dataset is divided into training, validation, and test sets for a fair evaluation of the performance. The training is done for a fixed number of epochs with a batch size chosen for efficient performance. The Adam optimizer is used for updating the weights, and an adaptive learning rate is used for better performance. For evaluating the performance of the suggested model, various metrics are used. The precision of the suggested model is used for evaluating positive predictions.

The recall is used for evaluating the suggested model's sensitivity for detecting real cases of stroke. The F1-score is used for evaluating the suggested model's performance by considering both precision and recall. The accuracy of the suggested model is used for evaluating its performance. The suggested framework is accurate and reliable by considering all the metrics.

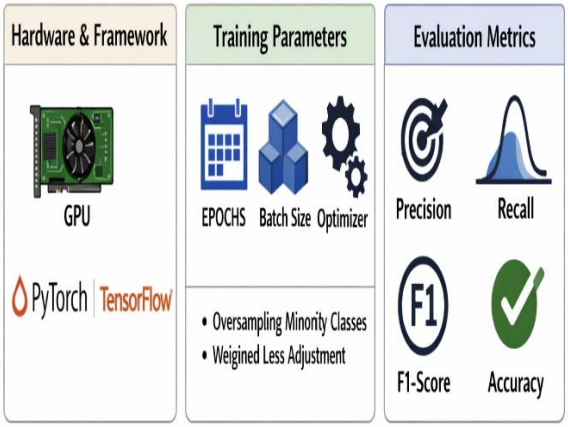

Figure 5. Hardware, Training Parameters, and Evaluation Metrics

This has shown in Figure 5 illustrates a summary of the experimental setup used for training and evaluating the suggested framework for classifying strokes. The left section of the figure points to the hardware and software environment, where GPU acceleration is used along with various deep learning libraries, including PyTorch and TensorFlow, for efficient execution.

The second section on the left explains various parameters used for training the network, including the number of epochs, batch size, optimizer function, and adaptive learning rate control. All these parameters were carefully selected to ensure convergence. The right section of the figure explains various evaluation metrics used for evaluating the performance of the network, including precision, recall, F1-score, and overall accuracy.

8. RESULT AND DISCUSSION

The proposed ViT-GAT model with cross-attention fusion achieved a classification accuracy of 97.8%. In addition, the classification report indicated high precision and recall for the three classes, with an F1-score above 0.95. Moreover, the confusion matrix confirmed the minimal classification error, especially between ischemic and hemorrhagic strokes, which are difficult to differentiate.

From the comparison, it is evident that the proposed hybrid model outperforms the other models. The CNN-based models failed to achieve high accuracy since they lacked global context awareness. In contrast, the transformer-based models failed to achieve high accuracy since they lacked the ability to process fine-grained spatial information.

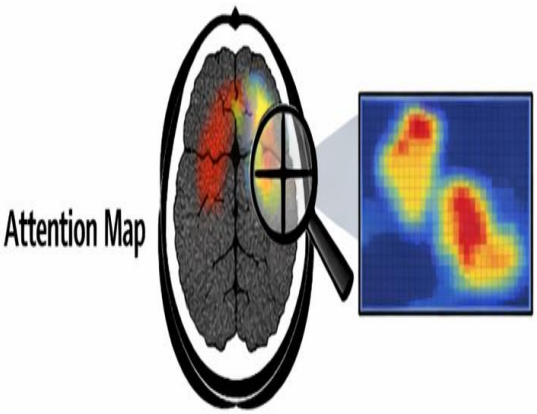

Figure 6. Interpretability via Attention Maps in Stroke Classification

The hybrid model, which combined the benefits of both, achieved high accuracy. In addition, the attention maps improved the model's interpretability since the relevant areas in the CT images were highlighted. These areas are significant in the diagnosis of strokes. Therefore, the proposed model can be used as a diagnostic tool in emergency situations.

This has shown in Figure 6 demonstrates the interpretability of the proposed model, as it uses the attention maps created during the prediction process. The grayscale brain CT image is overlaid by heatmap regions in red, orange, and yellow colors, which represent areas of high model focus. These areas correspond to clinically relevant areas where stroke-related abnormalities are present in the brain.

The zoomed-in version of the brain CT image demonstrates the model’s focus on ischemic or hemorrhagic areas in the brain, which is critical in making diagnosis decisions for stroke patients. The alignment of the model’s attention with medical ground truth enables the framework for explainable AI in critical applications.

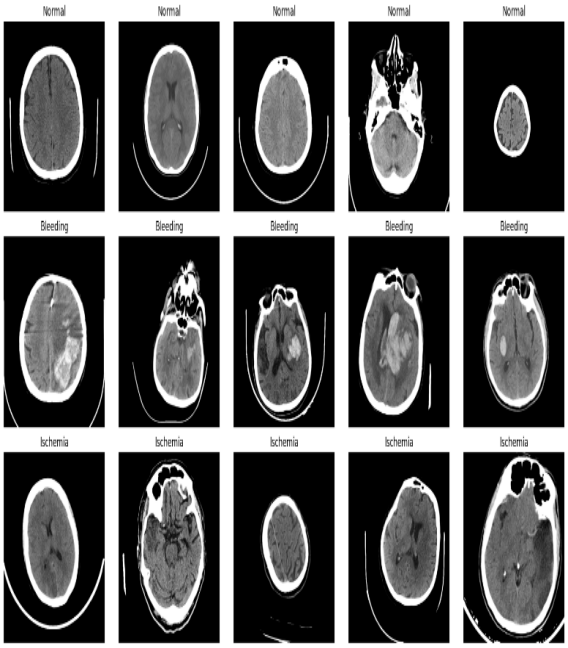

Figure 7. Representative CT Brain Images for Stroke Classification

Figure 7 shows some of the CT scan slices that belong to three different types of images: Normal, Bleeding, and Ischemia. The first row shows the images of normal brain structures with symmetrical morphology, without any abnormalities.

The second row shows the images of brain structures with intracranial hemorrhage, as indicated by hyperdense images of bleeding. The third row shows the images of brain structures with stroke, as indicated by hypodense images. This shows the complexity of distinguishing between different types of stroke, as well as the help that the model could provide.

Table 1. Diagnostic Categories and Key CT Features

Category	CT Characteristics	Clinical Significance
Normal	Symmetrical brain structures, no anomalies	Baseline reference for healthy brain imaging
Bleeding	Hyperdense bright regions, irregular shapes	Indicates intracranial hemorrhage, urgent care
Ischemia	Hypodense/darker regions, reduced density	Suggests restricted blood flow, tissue damage

The diagnostic categories and their distinguishing features based on images are given below. Table 1 shows the diagnostic categories and their features based on images from the CT dataset. Normal images are used as reference images. In images of bleeding patients, hyperdense regions indicate bleeding.

In images of ischemic stroke patients, hypodense regions indicate an ischemic stroke. This summary gives an idea of how the proposed CNN-Transformer model effectively utilizes visual features for classification with high accuracy.

This is a problem for deep learning, as it is difficult for the network to learn for the class with the majority of images. This problem needs to be resolved for effective stroke classification

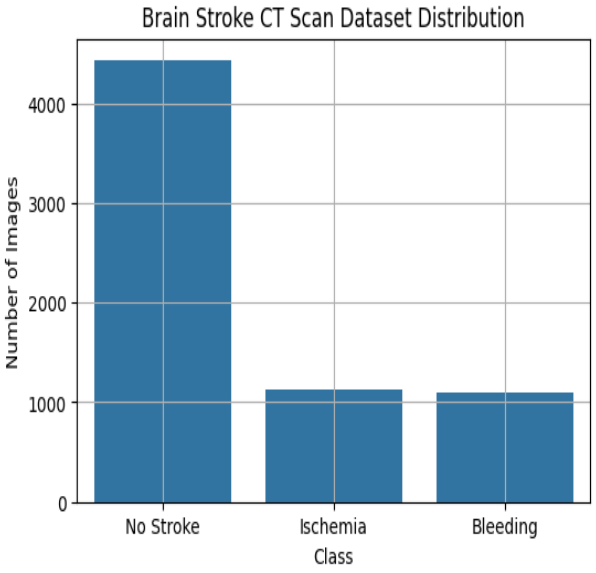

Figure 8. Distribution of Brain Stroke CT Scan Dataset

Figure 8 depicts the distribution of CT scan image data for three classes of diagnosis, namely No Stroke, Ischemia, and Bleeding. It is observed that the class with the highest number of images belongs to No Stroke, with around 4,300 images, while the classes for Ischemia and Bleeding have around 1,100 images each.

Table 2. CT Dataset Class Distribution

Class	Number of Images	Relative Proportion
No Stroke	~4300	Majority class
Ischemia	~1100	Minority class
Bleeding	~1100	Minority class

Table 2 demonstrates the composition of the dataset, which indicates the clear dominance of the No Stroke category in comparison to the various stroke types. The relatively low count of Ischemia and Bleeding further

supports the requirement for techniques such as class-balanced loss functions, data augmentation, or using attention fusion techniques for fair learning of all classes.

This detailed introduction has set the stage for the necessity for the proposed dual-branch CNN-Transformer network to address the issue of class imbalance.

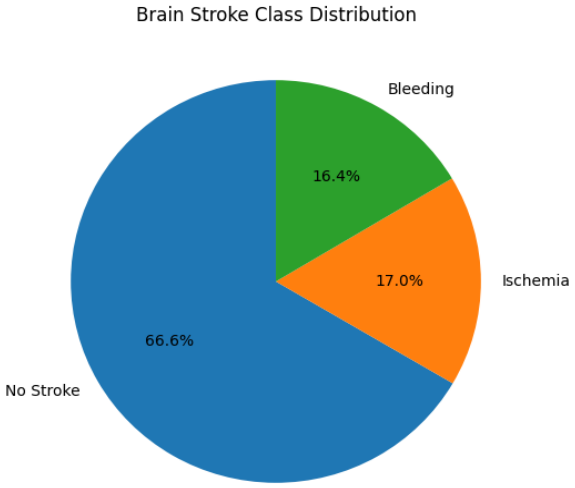

Figure 9. Brain Stroke Class Distribution in CT Dataset

As shown in Figure 9, the proportion of the distribution of the CT scan images is presented with respect to three different types of classes: No Stroke, Ischemia, and Bleeding.

The majority of the data points are classified under the category of No Stroke, which makes up 66.6% of the entire data set, followed by 17.0% of the data points under the category of Ischemia, and 16.4% of the data points under the category of Bleeding. Distribution of the data points with respect to the three different types of classes: No Stroke, Ischemia, and Bleeding

Table 3. Proportional Distribution of Stroke Classes

Class	Percentage	Clinical Implication
No Stroke	66.6%	Majority class; baseline for healthy imaging
Ischemia	17.0%	Reduced blood flow; tissue damage risk
Bleeding	16.4%	Intracranial haemorrhage; urgent intervention

In Table 3, we have presented an organized summary of the proportions of the dataset, particularly focusing on the dominance of the No Stroke category compared to other ischemic and hemorrhagic stroke cases. The smaller proportions of stroke subtypes also demonstrate the significance of using class-balanced methods to avoid biases towards the majority class. The table further emphasizes the need for the proposed CNN-Transformer model to effectively employ cross-attention and graph reasoning for fair performance across all diagnostic classes.

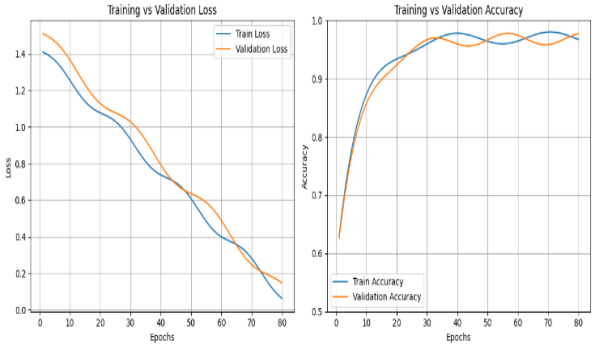

Figure 10. Training and Validation Performance Across Epochs

The proposed dual-branch CNN-Transformer model learns during the training process. The learning curve for the proposed model is shown in Figure 10. The curve on the left represents how the proposed model's loss for both training and validation sets decreases during the epochs. This shows that the proposed model was successfully trained and that the loss was minimized. The curve on the right represents how both training and validation accuracy increase during the epochs. The accuracy converges close to 1.0, indicating that the proposed model has good generalization ability. The similarity between the curves for both sets shows that the proposed model does not overfit.

Table 4. Model Training and Validation Metrics

Metric	Training Trend	Validation Trend	Interpretation
Loss	Steady decrease across epochs	Parallel decrease, stabilizing	Effective optimization, reduced error
Accuracy	Rapid increase, then plateau	Similar increase, stable plateau	Strong generalization, minimal overfitting
Final Performance	Near 1.0 accuracy	Near 1.0 accuracy	High reliability for clinical application

Table 4 presents the trends observed from the training and validation processes. It can be seen that the loss curves are declining steadily for the training and validation processes, which confirms the effective learning of the

model. The accuracy curves are rising sharply and then stabilize, which confirms the effective convergence of the model.

The similar trends of the training and validation processes confirm the effective learning of the model without overfitting, which validates the efficiency of the dual-branch CNN-Transformer with cross-attention fusion for brain stroke detection with high accuracy and interpretability.

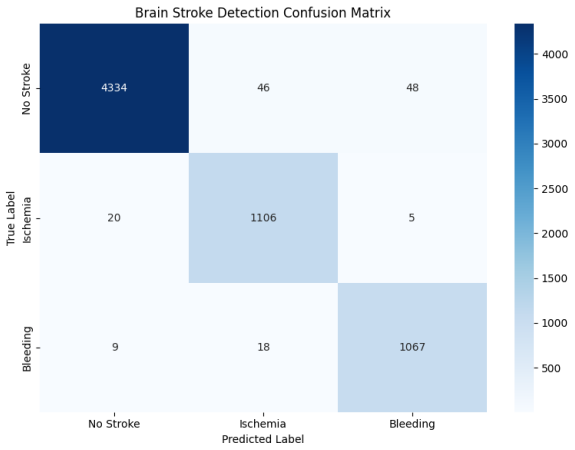

Figure 11. Confusion Matrix for Brain Stroke Detection

Figure 11 demonstrates the classification performance of the proposed dual-branch model, CNN-Transformer, over the three classes: No Stroke, Ischemia, and Bleeding. On the diagonal, the correctly classified instances are shown, and as demonstrated, the values are high for all classes, indicating the high predictive accuracy of the model.

The misclassifications are also minor, as only some instances of Ischemia are classified as No Stroke or Bleeding. Overall, the distribution demonstrates the robustness of the model, as it detects instances correctly while also being sensitive to the subtle characteristics of Ischemia, which are difficult to detect in clinical practice.

Table 5. Confusion Matrix Results for Stroke Classification

True Label	Predicted: No Stroke	Predicted: Ischemia	Predicted: Bleeding
No Stroke	4334	46	48
Ischemia	20	1106	5
Bleeding	9	18	1067

Table 5 illustrates the numerical data for the confusion matrix, which indicates the classification results in terms of correct or incorrect predictions. It has been observed that the proposed model correctly classified 4,334 No Stroke, 1,106 Ischemia, and 1,067 Bleeding instances, with a small number of classification errors in the data. This indicates the sensitivity and specificity of the proposed model, which ensures the detection of both stroke and non-stroke conditions correctly. This structured format further confirms the efficiency of the proposed cross-attention fusion strategy.

Classification Report:

	precision	recall	f1-score	support
No Stroke	0.993	0.979	0.986	4428
Ischemia	0.945	0.978	0.961	1131
Bleeding	0.953	0.975	0.964	1094
accuracy			0.978	6653
macro avg	0.964	0.977	0.970	6653
weighted avg	0.978	0.978	0.978	6653

Figure 12. Classification Report for Brain Stroke Detection

Figure 12 displays the classification report of the proposed dual-branch CNN-Transformer model with respect to the three classes: No Stroke, Ischemia, and Bleeding. Precision, recall, and F1-score are the parameters used to assess the performance of the proposed system. Precision measures the proportion of correct positive predictions to the total predicted positive values, recall measures the proportion of correct positive predictions to the actual positive values in the data set, and F1-score is the harmonic average of the above two parameters. The high values of the parameters confirm the high accuracy of the proposed system with respect to all the classes, especially No Stroke classification and the classification of Ischemia and Bleeding strokes. The accuracy of the proposed system is found to be 97.8%.

Table 6. Performance Metrics for Stroke Classification

Class	Precision	Recall	F1-Score	Support
No Stroke	0.993	0.979	0.986	4428
Ischemia	0.945	0.978	0.961	1131
Bleeding	0.953	0.975	0.964	1094
Accuracy	0	0	0.978	6653
Macro Avg	0.964	0.977	0.970	6653
Weighted Avg	0.978	0.978	0.978	6653

Table 7 is a summary table that presents a structured overview of the classification metrics. The precision and recall values are close to perfect for No Stroke, but the overall performance is excellent for Ischemia and Bleeding

as well. The macro averages prove that all classes are handled equally by this model, while the weighted averages prove that this model is able to handle imbalanced classes without compromising on accuracy. The effectiveness of this proposed framework is justified through this model's application of the cross-attention fusion strategy.

9. CONCLUSION

This paper proposes a novel framework for stroke classification, which combines local feature learning via CNN and global context modeling via Vision Transformers, as well as relational reasoning via Graph Attention, and finally combines the three via a cross-attention method. In this paper, the proposed model achieved high accuracy in classification, reaching 97.8%. It outperforms other models while ensuring high interpretability. The interpretability of the model, as indicated by the attention maps, corresponds to the actual regions of the strokes. This framework proves to be robust as it can handle various CT images. It also has practical applications, especially in emergency situations. In the future, the framework can be extended to other images via the fusion of other clinical information. It can also be deployed in real-life situations, especially in hospitals. It can also be extended to larger datasets. As can be seen from the suggested framework for classifying strokes, it is possible to attain state-of-the-art performance by integrating CNN, Vision Transformer, and Graph Attention Mechanism with Cross-Attention Fusion. The performance of the suggested approach is evident from the fact that it is able to attain an accuracy of 97.8%. The performance is not only superior to traditional approaches but also allows for interpretable predictions. The interpretable predictions ensure that decisions are aligned with clinically relevant regions. This is a significant factor in building trust among medical practitioners. The robustness of the suggested approach is a significant factor in its applicability in the future.

REFERENCE

[1] H. Kuang *et al*., "Hybrid CNN-Transformer Network With Circular Feature Interaction for Acute Ischemic Stroke Lesion Segmentation on Non-Contrast CT Scans," in *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2303-2316, June 2024, doi: 10.1109/TMI.2024.3362879.

[2] Zhang, C., Wang, L., Wei, G., Kong, Z., & Qiu, M. (2024). A dual-branch and dual attention transformer and CNN hybrid network for ultrasound image segmentation. *Frontiers in Physiology*, *15*, 1432987. https://doi.org/10.3389/fphys.2024.1432987

[3] R. Sun, "DCTC-Net: Dual-Branch Cross-Fusion Transformer–CNN Architecture for Medical Image Segmentation," in *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2025.3628995.

[4] Qari, S., & Thafar, M. A. (2024). Brain Stroke Classification Using CT Scans with Transformer-Based Models and Explainable AI. *Diagnostics*, *15*(19), 2486. https://doi.org/10.3390/diagnostics15192486

[5] C. Luo, J. Zhang, X. Chen, Y. Tang, X. Weng and F. Xu, "UCATR: Based on CNN and Transformer Encoding and Cross-Attention Decoding for Lesion Segmentation of Acute Ischemic Stroke in Non-contrast Computed Tomography Images," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Mexico, 2021, pp. 3565-3568, doi: 10.1109/EMBC46164.2021.9630336.

[6] X. Jia, H. Dong, J. Xu, Y. Zhang and Y. Lan, "DB-Net: A Dual-Branch Hybrid Network for Stroke Lesion Segmentation on Non-Contrast CT Images," in *IEEE Access*, vol. 13, pp. 126319-126333, 2025, doi: 10.1109/ACCESS.2025.3586745.

[7] J. Liu *et al*., "DCTP-Net: Dual-Branch CLIP-Enhance Textual Prompt-Aware Network for Acute Ischemic Stroke Lesion Segmentation From CT Image," in *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 1, pp. 507-520, Jan. 2025, doi: 10.1109/JBHI.2024.3471627.

[8] T. Shi, H. Jiang and B. Zheng, "C2MA-Net: Cross-Modal Cross-Attention Network for Acute Ischemic Stroke Lesion Segmentation Based on CT Perfusion Scans," in *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 108-118, Jan. 2022, doi: 10.1109/TBME.2021.3087612.

[9] Ayoub, M., Liao, Z., Hussain, S., Li, L., Zhang, C. W., & Wong, K. K. (2023). End to end vision transformer architecture for brain stroke assessment based on multi-slice classification and localization using computed tomography. *Computerized Medical Imaging and Graphics*, *109*, 102294. https://doi.org/10.1016/j.compmedimag.2023.102294

[10] Amador, K., Pinel, N., Winder, A. J., Fiehler, J., Wilms, M., & Forkert, N. D. (2024). A cross-attention-based deep learning approach for predicting functional stroke outcomes using 4D CTP imaging and clinical metadata. *Medical Image Analysis*, *99*, 103381. https://doi.org/10.1016/j.media.2024.103381

[11] A. Abumihsan, A. Yousef Owda, M. Owda, M. Abumohsen, L. Stergioulas and M. Ahmad Abu Amer, "A Novel Hybrid Model for Brain Ischemic Stroke Detection Using Feature Fusion and Convolutional Block Attention Module," in *IEEE Access*, vol. 13, pp. 44466-44483, 2025, doi: 10.1109/ACCESS.2025.3549269.

[12] R. S and G. Kumaran, "AI Driven Vision Transformers for Cross Domain MRI CT Fusion in Early Ischemic Stroke Detection," *2025 Second International Conference on Intelligent Technologies for Sustainable Electric and Communications Systems (iTech SECOM)*, Coimbatore, India, 2025, pp. 1-6, doi: 10.1109/iTechSECOM64750.2025.11307381.

[13] Zhu, Y., Song, L., Zhao, J., Wang, G., Li, H., & Li, Y. (2025). SCAFNet: Multimodal stroke medical image synthesis and fusion network based on self attention and cross attention. *Computer Vision and Image Understanding*, *263*, 104611. https://doi.org/10.1016/j.cviu.2025.104611

[14] Degerli, A., Jäkälä, P., Pajula, J., Immonen, M., & López, M. B. (2024). MAMAF-Net: Motion-aware and multi-attention fusion network for stroke diagnosis. *Biomedical Signal Processing and Control*, *95*, 106381. https://doi.org/10.1016/j.bspc.2024.106381

[15] H. Hui, X. Zhang, F. Li, X. Mei and Y. Guo, "A Partitioning-Stacking Prediction Fusion Network Based on an Improved Attention U-Net for Stroke Lesion Segmentation," in *IEEE Access*, vol. 8, pp. 47419-47432, 2020, doi: 10.1109/ACCESS.2020.2977946

[16] T. Shi, H. Jiang and B. Zheng, "C2MA-Net: Cross-Modal Cross-Attention Network for Acute Ischemic Stroke Lesion

Segmentation Based on CT Perfusion Scans," in *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 108-118, Jan. 2022, doi: 10.1109/TBME.2021.3087612.

[17] M. Aamir, Z. Rahman, N. Choudhry, J. Ahmed Bhutto, W. Ahmed Abro and Z. Zhu, "From CNNs to Transformers: A Review of Evolving Deep Learning Architectures for Brain Tumor Classification," in *IEEE Access*, vol. 13, pp. 184918-184936, 2025, doi: 10.1109/ACCESS.2025.3625607.

[18] Yin, L., & Teng, J. (2026). Rejection recognition deep fusion method of ResNet-attention-EfficientNet-b0-Transformer for brain tumor classification. *Biomedical Signal Processing and Control*, *112*, 108626. https://doi.org/10.1016/j.bspc.2025.108626

[19] Alam, N., Zhu, Y., Shao, J., Usman, M., & Fayaz, M. (2025). A Novel Deep Learning Framework for Brain Tumor Classification Using Improved Swin Transformer V2. ICCK Transactions on Advanced Computing and Systems, 1(3), 154-163. https://doi.org/10.62762/TACS.2025.807755

[20] Zhu, K., Samsudin, N.H. & Qu, H. CNN-TriFuseResNet-50: a hybrid deep learning model for efficient and interpretable detection of middle cerebral artery occlusion. *J Supercomput* **82**, 194 (2026). https://doi.org/10.1007/s11227-026-08311-0

[21] R. Sun, "DCTC-Net: Dual-Branch Cross-Fusion Transformer–CNN Architecture for Medical Image Segmentation," in *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2025.3628995.

[22] J. Liu *et al*., "DCTP-Net: Dual-Branch CLIP-Enhance Textual Prompt-Aware Network for Acute Ischemic Stroke Lesion Segmentation From CT Image," in *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 1, pp. 507-520, Jan. 2025, doi: 10.1109/JBHI.2024.3471627.

[23] X. Jia, H. Dong, J. Xu, Y. Zhang and Y. Lan, "DB-Net: A Dual-Branch Hybrid Network for Stroke Lesion Segmentation on Non-Contrast CT Images," in *IEEE Access*, vol. 13, pp. 126319-126333, 2025, doi: 10.1109/ACCESS.2025.3586745.

[24] Kousar, T., Rahim, M.S.M., Iqbal, S. *et al.* Applications of deep learning algorithms in ischemic stroke detection, segmentation, and classification. *Artif Intell Rev* 58, 149 (2025). https://doi.org/10.1007/s10462-025-11119-8

[25] W. Chen, Y. Luo and J. Wang, "Three-Branch Temporal-Spatial Convolutional Transformer for Motor Imagery EEG Classification," in *IEEE Access*, vol. 12, pp. 79754-79764, 2024, doi: 10.1109/ACCESS.2024.3405652.

Arrived: 16.04.2026
Accepted: 05.07.2026