2026 1st International Conference on Next-Generation Electrical & Electronics, Computer Systems, and Technologies (iCONEECT)
25–26 September, 2026, Premier University, Chittagong, Bangladesh

A Gated Three-Stream Fusion Framework for Automated Brain-Stroke Classification from MRI

1st Mugdha Farheen Kibria
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
mugdhafarheenkibria@gmail.com

2nd Zarin Tasnim Vega
Department of Computer Science
BRAC University
Dhaka-1212, Bangladesh
zarintasnimvega@gmail.com

3rd Ahmad Farruk Towsif
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
towsif2000@gmail.com

4th Md. Samiul Islam
Department of Computer Science
BRAC University
Dhaka-1212, Bangladesh
samiul.sakib111@gmail.com

5th Azizul Kabir Jayed
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
azizulkabirjayed@gmail.com

6th N M Saif Kabir
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
saifkabeer14@gmail.com

7th Md. Abdus Sium
Department of Electrical and Electronic Engineering
BRAC University
Dhaka-1212, Bangladesh
siumprof9005@gmail.com

8th Tauhidur Rahman Sakib*
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
tauhidurrahmansakib@gmail.com

***Abstract*—Fast and correct separation of haemorrhagic and ischemic stroke from brain MRI is critical for emergency care, but it is hard because lesion appearance varies and subtype cues are subtle. We present TriFusion-ResViTNet, a three-stream framework that learns ResNet-50 spatial features, Vision Transformer (ViT) global context, and radiomics texture-frequency descriptors, and merges them with a gated-attention module that weights each stream by reliability before a dense softmax classifier. The model was trained on a primary cohort of 4,554 scans (haemorrhagic, ischemic, normal) and tested on an independent secondary cohort of 1,430 scans. On the primary test set it reached 98.57% accuracy, a macro F1-score of 0.985, and a macro ROC-AUC of 0.993, beating single-stream and hybrid baselines, and it held 98.25% accuracy externally. Grad-CAM with edge, contour, and intensity analyses confirmed anatomically plausible attention, while the Cohen kappa coefficient, the McNemar test, and confidence intervals supported reliability.**

***Index Terms*—stroke classification, magnetic resonance imaging, vision transformer, ResNet-50, radiomics, gated-attention fusion, explainable artificial intelligence.**

I. Introduction

Stroke is among the foremost causes of death and long-term disability worldwide, and in 2019 it ranked as the second-leading cause of death globally [1]. Clinically, it divides into two dominant subtypes that demand opposite first-line treatments: ischemic stroke, caused by occlusion of a cerebral vessel, and haemorrhagic stroke, caused by vessel rupture and bleeding into brain tissue. Because thrombolytic therapy that can rescue ischemic tissue is dangerous in the presence of haemorrhage, an incorrect or delayed subtype decision carries severe consequences, and the available treatment window is narrow. Magnetic resonance imaging (MRI), and diffusion-weighted imaging in particular, resolves early ischemic and haemorrhagic changes with high sensitivity, yet expert interpretation is time-consuming and shows non-trivial inter-reader variability, especially for subtle or early lesions. High stakes, tight timing, and limited specialist availability together motivate automated, reproducible decision support for MRI-based stroke classification.

Early brain-MRI classifiers used handcrafted intensity and texture features, whereas CNNs learn hierarchical representations. Common backbones include VGG [2], ResNet [3], DenseNet [4], and EfficientNet [5]; Khalaf et al. combined EfficientNet with metaheuristic optimisation for three-class stroke classification [6]. CNNs capture local structure but have limited long-range context. Self-attention [7], introduced to vision through ViT [8], models global dependencies, although its data requirements motivate hybrid CNN-Transformer designs for small medical datasets. TransMed combines convolutional encoders and Transformer reasoning [9]; residual ViT uses self-supervised pre-training for brain-tumour MRI [10]; other hybrids use cross-attention [11] or modality-correlated

979-8-3195-4810-8/26/$31.00 ©2026 IEEE

cross-attention [12]. For stroke, ViT with a bidirectional LSTM reported strong subtype discrimination [13]. Radiomics provides complementary texture-frequency information: gray-level co-occurrence matrix (GLCM) and wavelet descriptors are discriminative on brain MRI [14], and radiomic–deep feature fusion improves classification [15]. Grad-CAM supports model interpretation [16]. Outside medical imaging, Chowdhury et al. explored hybrid quantum-classical fruit classification and calorie prediction [17], while Mazumder et al. studied machine-learning speaker identification [18]. These cross-domain applications do not directly validate MRI stroke classification. Few stroke models combine CNN, Transformer, and radiomics streams with adaptive fusion, explainability, and external validation; this work targets that gap.

This paper proposes TriFusion-ResViTNet, a three-stream framework that learns ResNet-50 spatial features, ViT global context, and radiomics texture-frequency descriptors, and fuses them through a gated-attention module that weights each stream by its reliability before classification. We evaluate it on a primary cohort and an independent secondary cohort for three-way classification of haemorrhagic, ischemic, and normal scans. The main contributions are as follows.

- A three-stream fusion network that unifies CNN, Transformer, and radiomics representations for brain-MRI stroke classification.
- A gated-attention fusion module with multi-scale refinement that adaptively weights the streams and improves on naive concatenation.
- A full study with ablation, comparison against traditional and hybrid baselines, and external validation on a secondary cohort.
- An interpretability and reliability assessment using Grad-CAM with edge, contour, and intensity analyses, plus the Cohen kappa coefficient, the McNemar test, and confidence intervals.

Sections II–IV present the methodology, results, and conclusions.

II. Methodology

A. Dataset Description

Both cohorts are de-identified axial brain-MRI collections drawn from publicly released brain-stroke MRI repositories, with class labels provided by the source curators. The primary cohort contains 4,554 scans across three categories: haemorrhagic (1,457), ischemic (1,341), and normal (1,756). An independent secondary cohort of 1,430 scans (haemorrhagic 458, ischemic 421, normal 551) was reserved for external evaluation; it originates from a different acquisition setting, differing in scanner vendor and field strength, site, and sequence protocol, while covering the same three clinical categories. Both collections are anonymised public resources and were used in accordance with their respective data-use licences and ethical terms; no directly identifiable patient information was accessed, so this secondary analysis of already de-identified images required no additional institutional review board approval.

Each cohort used a stratified 70:10:20 train–validation–test split that preserved class proportions. Partitioning was performed at the patient level and before augmentation, so all slices belonging to one patient fall in a single subset and no patient contributes to more than one partition. Validation and test partitions were held out during fitting and model selection. Table I summarizes the splits.

TABLE I
Dataset Composition and Stratified 70/10/20 Splits

Dataset	Class	Total	Train	Val	Test
Primary	Haemorrhagic	1,457	1,020	146	291
	Ischemic	1,341	939	134	268
	Normal	1,756	1,229	176	351
	Subtotal	4,554	3,188	456	910
Secondary	Haemorrhagic	458	321	45	92
	Ischemic	421	295	42	84
	Normal	551	386	55	110
	Subtotal	1,430	1,002	142	286

B. Preprocessing and Augmentation

All scans were resized to 224×224 px and, because the source images are single-channel, replicated across three channels to match the input expected by the ImageNet-pretrained backbones. Pixel intensities were then rescaled to the unit interval, as in (1), where $I(x, y, c)$ denotes the original intensity at spatial location x, y in channel c.

$$I_{\text{norm}}(x, y, c) = \frac{I(x, y, c)}{255} \tag{1}$$

To improve generalization and limit overfitting, stochastic augmentation (random rotation, horizontal flipping, zoom, and width/height shifts) was applied to the training partition only; validation and test images underwent resizing and normalization alone. Visually near-duplicate slices were discarded before splitting: slices of the same class were compared pairwise with a perceptual similarity score that combines the structural similarity index with a normalised pixel-level mean-absolute-difference term, and one member of every pair above a fixed similarity threshold of 0.95 was removed. Because deduplication and patient-level partitioning both precede augmentation, near-identical anatomical slices cannot be distributed across subsets and augmented variants never cross partition boundaries, which removes the main routes to optimistic leakage.

C. Proposed TriFusion-ResViTNet Architecture

As shown in Fig. 1, TriFusion-ResViTNet processes each scan through three parallel encoders (a ResNet-50 spatial branch, a ViT contextual branch, and a radiomics texture-frequency branch) whose outputs are projected to a common dimensionality, fused by gated attention, refined, and classified. The design goal is to jointly capture local anatomical detail, global structural dependencies, and quantitative texture signatures within one end-to-end pipeline.

Spatial branch. The ResNet-50 encoder applies residual learning, in which each block adds a learned residual mapping to its input as in (2); global average pooling of the final feature maps yields a 2048-dimensional spatial descriptor, (3).

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{2}$$

$$f_C = \mathrm{GAP}(\phi_{\mathrm{ResNet}}(X)) \in \mathbb{R}^{2048} \tag{3}$$

Contextual branch. The ViT encoder splits the image into non-overlapping patches, linearly embeds them, prepends a class token, and adds positional encodings, (4). Stacked layers apply multi-head self-attention, (5), each combining attention and a feed-forward block with residual connections and layer normalisation, (6) and (7). The normalised class-token embedding of the final layer gives a 768-dimensional global descriptor, (8).

$$z_0 = [x_{cls};\, x_p^1 E;\, \cdots;\, x_p^N E] + E_{pos} \tag{4}$$

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\Big(QK^\top/\sqrt{d_k}\Big)\,V \tag{5}$$

$$z'_\ell = \mathrm{MSA}(\mathrm{LN}(z_{\ell-1})) + z_{\ell-1} \tag{6}$$

$$z_\ell = \mathrm{MLP}(\mathrm{LN}(z'_\ell)) + z'_\ell \tag{7}$$

$$f_V = \mathrm{LN}(z_L^0) \in \mathbb{R}^{768} \tag{8}$$

Texture-frequency branch. In parallel, handcrafted descriptors are extracted from the same scan. GLCM Haralick measures quantify second-order texture; contrast, energy, and homogeneity are defined in (9) and (10), where $p(i,j)$ is the normalised co-occurrence probability of gray levels i and j. Multi-scale frequency content is summarised by wavelet sub-band energies, (11). The concatenated descriptor r is passed through a dense embedding, (12), giving a 256-dimensional radiomics vector.

$$\mathit{Contrast} = \sum_{i,j} (i-j)^2\, p(i,j) \tag{9}$$

$$\mathit{Energy} = \sum_{i,j} p(i,j)^2, \quad \mathit{Homog} = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \tag{10}$$

$$E_b = \frac{1}{MN} \sum_{m,n} |W_b(m,n)|^2 \tag{11}$$

$$f_R = \sigma(W_r r + b_r) \in \mathbb{R}^{256} \tag{12}$$

Gated-attention fusion. The three descriptors are projected to a shared space, (13). A gating network assigns each branch a normalised reliability weight via a softmax over learned scores, (14), and the weighted descriptors are concatenated into the fused representation, (15). This adaptive weighting lets the network emphasise the most informative stream per sample, the key advantage over plain concatenation.

$$\tilde{f}_k = \mathrm{ReLU}(W_k f_k + b_k), \quad k \in \{C,V,R\} \tag{13}$$

$$g_k = \frac{\exp(w_k^\top f_k)}{\sum_j \exp(w_j^\top f_j)} \tag{14}$$

$$f_{fus} = \big\|_{k\in\{C,V,R\}}\, g_k \tilde{f}_k \tag{15}$$

After multi-scale refinement, the fused vector passes through fully connected layers with ReLU activation and dropout, ending in a softmax over the three classes, (16).

$$\hat{y} = \mathrm{softmax}(W_2\, \delta(W_1 f_{fus} + b_1) + b_2) \tag{16}$$

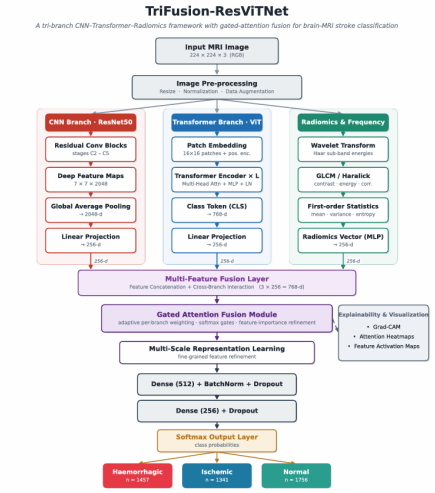


Fig. 1. Architecture of the proposed TriFusion-ResViTNet framework with gated-attention fusion of ResNet-50, ViT, and radiomics branches.

D. Loss Function and Training Setup

The network was optimised with the categorical cross-entropy objective (17), where y_i and the predicted probability denote class i and C equals 3. Optimisation used the Adam optimiser with dropout and training-only augmentation; the checkpoint with the best validation accuracy was retained and evaluated once on the held-out test set. All experiments were implemented in PyTorch and run on a single NVIDIA RTX 3090 (24 GB) GPU. For a fair comparison, every baseline and ablation variant shared the same split, preprocessing, optimizer, loss, and metrics.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{17}$$

E. Evaluation Metrics

Performance was quantified with accuracy, precision, recall, and F1-score, (18) to (20), computed per class and averaged, together with the ROC-AUC and the confusion matrix. Agreement beyond chance was measured with the Cohen kappa coefficient, (21), where p_o is the observed agreement and p_e the agreement expected by chance. The McNemar test and 95% confidence intervals were additionally used to assess significance and stability.

$$\mathit{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{18}$$

$$\mathit{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathit{Rec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{19}$$

$$F1 = 2 \cdot \frac{\mathit{Prec} \cdot \mathit{Rec}}{\mathit{Prec} + \mathit{Rec}} \tag{20}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{21}$$

III. Results and Discussion

This section reports overall and per-class performance on the primary test set, confusion matrices, ROC analysis, and training behaviour, then an ablation that isolates each branch

and fusion component, a benchmark against single-stream and hybrid baselines, external validation on the secondary cohort, qualitative interpretation through Grad-CAM and feature maps, and the computational and statistical analyses.

A. Overall Classification Performance

On the held-out primary test set (910 scans), TriFusion-ResViTNet reached an overall accuracy of 98.57% (897 of 910 scans classified correctly), with macro-averaged precision, recall, and F1-score of 0.985 and a macro ROC-AUC of 0.993, indicating balanced and discriminative behaviour across all three categories. Table II lists the per-class scores; the normal class was recognised most reliably, while the small residual error concentrated between the two stroke subtypes, which is expected given their occasionally overlapping appearance.

TABLE II
PER-CLASS PERFORMANCE ON THE PRIMARY TEST SET

Class	Precision	Recall	F1-score	Support
Haemorrhagic	0.9862	0.9828	0.9845	291
Ischemic	0.9777	0.9813	0.9795	268
Normal	0.9915	0.9915	0.9915	351
Macro average	0.9851	0.9852	0.9852	910
Weighted average	0.9857	0.9857	0.9857	910
Accuracy	—	—	0.9857	910

B. Confusion Matrix Analysis

Fig. 2 presents the confusion matrices for the primary and secondary test sets. The dominant diagonals confirm that the great majority of scans were assigned to their correct category, with only a handful of off-diagonal errors, chiefly between haemorrhagic and ischemic cases that share structural characteristics. The pattern is consistent across both cohorts, suggesting that the fusion strategy yields stable class-wise behaviour rather than over-fitting a single distribution.

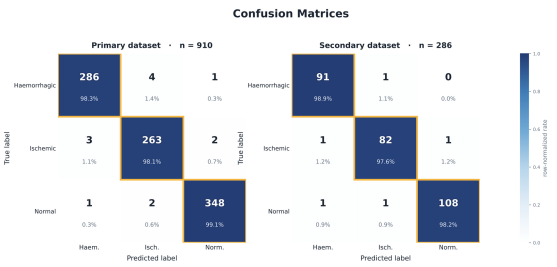


Fig. 2. Confusion matrices on the primary ($n = 910$) and secondary ($n = 286$) test sets; diagonal cells denote correct predictions.

C. ROC Curve Analysis

The one-vs-rest ROC analysis in Fig. 3 corroborates the strong separability of the three classes, with per-class areas under the curve approaching the macro value of 0.993. High true-positive rates are kept at low false-positive rates, implying that the complementary spatial, contextual, and texture cues jointly sharpen the decision boundaries between similar lesions.

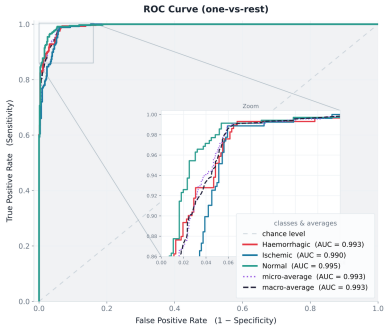


Fig. 3. One-vs-rest ROC curves with micro- and macro-averages on the primary test set.

D. Training Convergence

Fig. 4 shows the learning curves. Training and validation accuracy rise in tandem while the corresponding losses decline, and the modest gap between the two curves, together with the early-stopping checkpoint, indicates effective regularisation by dropout and training-only augmentation without pronounced overfitting.

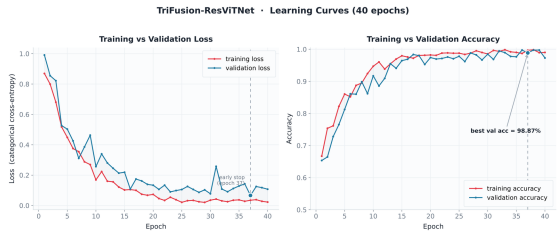


Fig. 4. Training and validation accuracy and loss over the training schedule.

E. Ablation Study

To isolate the contribution of each component, Table III evaluates standalone branches, two-branch combinations, and fusion variants under an identical protocol. Performance increases as branches are added, and replacing plain concatenation with gated attention, then adding multi-scale refinement, yields further gains. Removing the radiomics branch is the single most damaging ablation, underlining the value of texture-frequency cues that the learned branches do not fully capture.

TABLE III
ABLATION STUDY ON THE PRIMARY TEST SET

Configuration	Branches	Fusion	Acc. (%)	F1	AUC
Radiomics only	R	—	88.24	0.8796	0.9503
ViT only	V	—	93.96	0.9380	0.9791
ResNet-50 only	C	—	94.62	0.9444	0.9806
ViT + Radiomics	V+R	gated	95.93	0.9586	0.9840
CNN + Radiomics	C+R	gated	96.59	0.9646	0.9876
CNN + ViT	C+V	gated	97.14	0.9701	0.9887
Tri-branch	C+V+R	concat	97.58	0.9748	0.9899
Tri-branch minus multi-scale	C+V+R	gated	98.13	0.9805	0.9914
Full TriFusion-ResViTNet	C+V+R	gated+MS	**98.57**	**0.9852**	**0.9926**

F. Comparison with Baseline and Hybrid Models

Table IV compares the proposed framework against widely used single-stream architectures and representative hybrid designs under the same data and training conditions. Traditional CNN backbones cluster in the low-to-mid nineties, the standalone ViT is comparable, and dual-branch hybrids improve further; TriFusion-ResViTNet attains the best score on every metric. The largest margin appears over the strongest competing hybrid, which we attribute to the radiomics stream and the reliability-aware gated fusion.

TABLE IV
COMPARISON WITH TRADITIONAL AND HYBRID MODELS

Model	Type	Acc.	Prec.	Rec.	F1	AUC
VGG16 [2]	CNN	91.76	0.9168	0.9142	0.9155	0.9711
ViT-B/16 [8]	Transformer	93.96	0.9388	0.9372	0.9380	0.9791
ResNet-50 [3]	CNN	94.62	0.9451	0.9438	0.9444	0.9806
DenseNet-121 [4]	CNN	95.05	0.9498	0.9486	0.9492	0.9827
EfficientNet-B0 [5]	CNN	95.38	0.9531	0.9520	0.9525	0.9843
CNN + BiLSTM [13]	Hybrid	95.82	0.9575	0.9563	0.9569	0.9861
ResNet-50 + Radiomics [15]	Hybrid	96.59	0.9652	0.9641	0.9646	0.9876
ResNet-50 + ViT [10]	Hybrid	96.92	0.9686	0.9675	0.9680	0.9891
CNN + Transformer (cross-attn.) [11]	Hybrid	97.36	0.9731	0.9722	0.9726	0.9906
TriFusion-ResViTNet	Hybrid	**98.57**	**0.9851**	**0.9852**	**0.9852**	**0.9926**

G. Cross-Dataset Generalisation

Robustness was assessed on the independent secondary cohort, which differs from the primary set in scanner vendor and field strength, imaging site, and sequence protocol, and therefore in contrast rendering, slice appearance, and noise characteristics. As reported in Table V, every metric degraded by less than half a percentage point, accuracy falling from 98.57% to 98.25% (281 of 286 scans correct), macro F1 from 0.9852 to 0.9821, and macro AUC from 0.9926 to 0.9904.

Two properties of the design plausibly explain so small a drop. The cues that survive a protocol change—lesion geometry, hemispheric asymmetry, and relative texture contrast—are those the spatial and radiomics branches rely on, whereas absolute intensity calibration, the quantity a vendor change affects most, is largely removed by per-image normalisation. The gating weights are also computed per sample, so a stream that becomes unreliable under shifted acquisition statistics is down-weighted rather than contributing a fixed share. The margin is nonetheless narrow and needs confirmation on further cohorts. Fig. 5 shows representative batch predictions on unseen scans, where predicted labels match the references with high confidence.

TABLE V
CROSS-DATASET GENERALISATION: PRIMARY VS. SECONDARY

Metric	Primary (n=910)	Secondary (n=286)	Diff.
Accuracy	98.57%	98.25%	−0.32
Macro precision	0.9851	0.9818	−0.0033
Macro recall	0.9852	0.9824	−0.0028
Macro F1	0.9852	0.9821	−0.0031
Weighted F1	0.9857	0.9825	−0.0032
Cohen kappa	0.9784	0.9736	−0.0048
Macro AUC	0.9926	0.9904	−0.0022

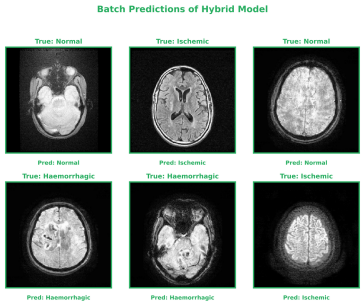


Fig. 5. Representative batch predictions on unseen secondary-cohort scans, with predicted class and confidence.

H. Qualitative Interpretation and Feature Analysis

To examine what the network attends to, Grad-CAM heatmaps were combined with edge detection, contour extraction, and intensity profiling. As in Fig. 6, the activation peaks coincide with anatomically meaningful regions, the extracted contours trace lesion and structural boundaries, and the intensity profile peaks at the highlighted region, giving a coherent, human-readable account of each decision. Fig. 7 visualises intermediate feature maps, which progress from generic edges and textures in early layers to increasingly organised, class-relevant patterns in deeper layers.

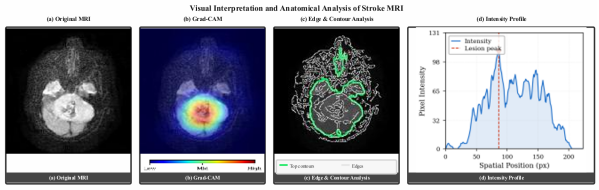


Fig. 6. Visual interpretation using Grad-CAM, edge detection, contour analysis, and intensity profiling for a representative stroke scan.

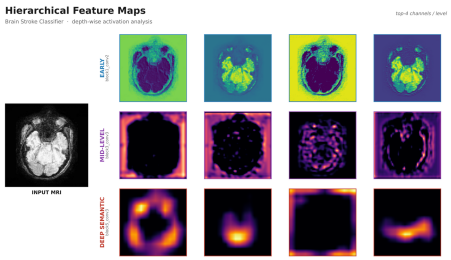


Fig. 7. Feature-map visualisation across network depth, from low-level texture to organised, discriminative patterns.

I. Statistical Validation and Computational Cost

Several statistical checks gauged reliability, summarised in the upper part of Table VI. A Cohen kappa of 0.9784 reflects almost-perfect agreement between predictions and ground truth; the McNemar test indicates that the improvement over the strongest baseline is statistically significant; and the 95% confidence interval for accuracy is narrow, denoting stable behaviour across the evaluation sample.

Because emergency-room deployment depends on the cost of a three-stream design, complexity and speed were measured on the implemented model rather than estimated. The network holds 112.3 M trainable parameters and requires 22.1 GFLOPs for one forward pass on a 224×224 scan, counting the ResNet-50, ViT-B/16, radiomics embedding, gated-attention fusion, and classifier. Mean inference latency on one NVIDIA RTX 3090 is 31.4 ms per scan, about 31.8 scans per second, excluding data loading and the offline radiomics pre-computation. Latency is thus negligible beside the acquisition and reading time of an acute stroke study, although the parameter count still argues for compression before deployment on commodity or edge hardware.

TABLE VI
STATISTICAL VALIDATION AND MEASURED COMPUTATIONAL COST

Statistical analysis	**Result**	**Interpretation**
Cohen kappa	0.9784	Almost-perfect agreement
McNemar test	$p = 0.013$	Significant vs. best baseline
95% CI (accuracy)	97.8% – 99.1%	Stable performance
Trainable parameters	112.3 M	Measured on implementation
Complexity per scan	22.1 GFLOPs	224×224 forward pass
Inference latency	31.4 ms	RTX 3090, $\approx$ 31.8 scans/s

IV. Conclusion

TriFusion-ResViTNet combines ResNet-50, ViT, and radiomics through reliability-aware gated attention for MRI stroke classification. Primary-cohort accuracy was 98.57%, macro F1 was 0.985, and macro ROC-AUC was 0.993, exceeding single-stream and hybrid baselines with only a slight decline on the secondary cohort. Ablations, Grad-CAM, and statistical tests support the contributions of the branches and fusion module and the plausibility of predictions. Measured cost was 112.3 M parameters, 22.1 GFLOPs, and 31.4 ms per scan on an RTX 3090. Limitations include 2D inputs, limited data sources, high edge-deployment cost, and no distinction between intracerebral and subarachnoid haemorrhage. Future work will investigate volumetric multi-sequence MRI, larger multi-centre cohorts with clinically confirmed labels, finer subtypes, model compression, and privacy-preserving federated training.

Author Contribution

N M Saif Kabir supervised the study. Tauhidur Rahman Sakib designed the architecture and serves as corresponding author. The other six authors contributed equally to all remaining research tasks and manuscript preparation.

Acknowledgment

The authors used a large language model to assist with grammar correction and language refinement. All technical content, results, and conclusions are the authors' own, and the authors reviewed and take full responsibility for the final manuscript.

References

[1] V. L. Feigin et al., "Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," The Lancet Neurology, vol. 20, no. 10, pp. 795–820, Sep. 2021.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, Sep. 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770–778.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. CVPR, 2017, pp. 2261–2269.

[5] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," arXiv:1905.11946, May 2019.

[6] X. Xue, S. E. Viswapriya, D. Rajeswari, R. Z. Homod, and O. I. Khalaf, "An efficient classification framework for brain stroke prediction using shuffled shepherd optimization," Scientific Reports, vol. 15, no. 1, p. 33516, Sep. 2025.

[7] A. Vaswani et al., "Attention is all you need," arXiv:1706.03762, Jun. 2017.

[8] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv:2010.11929, Oct. 2020.

[9] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multi-modal medical image classification," Diagnostics, vol. 11, no. 8, p. 1384, Jul. 2021.

[10] M. A. Karagoz, O. U. Nalbantoglu, and G. C. Fox, "Residual vision transformer (ResViT) based self-supervised learning model for brain tumor classification," arXiv:2411.12874, Nov. 2024.

[11] G. Jayaraman, S. Meganathan, S. S. M. Shah, M. Anuradha, R. K. Sundararajan, and R. Rajakumar, "A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification," Scientific Reports, vol. 15, no. 1, p. 45769, Nov. 2025.

[12] J. Lin et al., "CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation," IEEE Transactions on Medical Imaging, vol. 42, no. 8, pp. 2451–2461, Feb. 2023.

[13] R. Samuel and T. Pandi, "MRI-based brain stroke classification using a hybrid vision transformer–BiLSTM architecture," Frontiers in Neurology, vol. 17, p. 1802972, Apr. 2026.

[14] M. Jafari et al., "Hybrid radiomics and machine learning for brain tumor multi-task classification: an exploratory study integrating GLCM and convolution-based features of shape and intensity," Health Science Reports, vol. 8, no. 8, p. e71195, Aug. 2025.

[15] L. Yin and J. Wang, "Enhancing brain tumor classification by integrating radiomics and deep learning features: a comprehensive study utilizing ensemble methods on MRI scans," Journal of X-Ray Science and Technology, vol. 33, no. 1, pp. 47–57, Dec. 2024.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. ICCV, 2017, pp. 618–626.

[17] M. T. Chowdhury et al., "A hybrid quantum-classical approach for fruit classification and calorie prediction using machine learning," in Quantum Shield for AI Security, A. Ali et al., Eds. Berlin, Germany: De Gruyter, 2026, ch. 12, pp. 273–298.

[18] S. I. Mazumder et al., "Smart voice signature: A machine learning approach to speaker identification," in Proc. ICCCT, Chennai, India, 2025, pp. 1–6.