Benchmarking Deep Vision Architectures: From Controlled Datasets to Real-World Clinical Validation

Raghda Essam Ali, Noha Ahmed Saad El-Dien, Magi Hossam Eldin Mahfouz
School of Computing and Digital Tech., Eslsca University, Egypt

***Abstract*—Stroke is a leading cause of mortality and long-term disability, making rapid and reliable detection from non-contrast computed tomography (CT) scans essential for timely clinical intervention. This study introduces NeuroCT-Bench, a unified and reproducible benchmark for evaluating deep learning architectures for automated stroke classification from brain CT images. The benchmark systematically compares five Convolutional Neural Networks (VGG16, ResNet50, DenseNet121, EfficientNetB0, MobileNetV2) and four Vision Transformers (Swin Transformer, ViT, DeiT, PVT-Small) under identical preprocessing, augmentation, and evaluation protocols using the Brain Stroke CT Image dataset from Kaggle, comprising 1,551 normal and 950 stroke slices. Internal validation using a deterministic 80/20 stratified split (seed = 42) demonstrated near-perfect performance for Transformer-based models, with PVT-Small and Swin Transformer achieving ROC-AUCs of 99.96% and 99.98%, respectively, while DenseNet121, VGG16, and EfficientNetB0 achieved strong CNN baselines with ROC-AUCs of 99.93%, 99.67%, and 99.05%. To evaluate robustness under real-world domain shift, the top-performing models were further assessed on an external patient-level clinical dataset containing 530 CT studies. EfficientNetB0 demonstrated the strongest generalization capability (accuracy: 76.92%, precision: 83.85%, ROC-AUC = 88.0%), whereas high-capacity Transformer models exhibited substantially larger performance degradation (ROC-AUCs of 75.0% for PVT-Small and 67.0% for Swin Transformer). These findings highlight the discrepancy between curated public datasets and heterogeneous clinical imaging conditions, emphasizing that high internal performance does not necessarily guarantee clinical robustness. In addition, an ablation study was conducted to evaluate a lightweight CNN–Transformer gated fusion strategy. Results demonstrate that adaptive fusion improves robustness and generalization compared with individual CNN or Transformer models and static feature concatenation. Overall, NeuroCT-Bench provides a transparent and reproducible framework for evaluating deep learning models for stroke analysis and supports future development of clinically deployable hybrid CNN–Transformer systems.**

***Keywords*—Stroke detection; brain CT image; deep learning; Convolutional Neural Networks; Vision Transformers; clinical validation; neuroimaging AI**

I. Introduction

Stroke remains a major global health concern, ranking among the leading causes of mortality and long-term disability and placing substantial pressure on healthcare systems worldwide [1], [2]. Rapid and accurate diagnosis is critical, particularly in emergency settings where treatment decisions must be made within minutes. Non-Contrast Computed Tomography (NCCT) is widely adopted as the first-line imaging modality due to its availability and speed; however, subtle early ischemic changes are often difficult to identify, and interpretation may vary across radiologists [3], [4]. These challenges have motivated the integration of artificial intelligence (AI) techniques into stroke assessment workflows, aiming to enhance diagnostic consistency and support clinical decision-making.

Traditional machine learning approaches—such as logistic regression, Support Vector Machines, and ensemble methods—offer a degree of interpretability but rely heavily on handcrafted feature engineering and often struggle with heterogeneous imaging conditions and patient variability [5], [6]. Convolutional Neural Networks (CNNs) significantly improved performance by learning hierarchical feature representations directly from imaging data [7], [8]. Nevertheless, CNNs are inherently constrained by local receptive fields, limiting their ability to model long-range spatial dependencies in medical images [9], [10].

Transformer-based architectures, originally introduced for natural language processing, address this limitation through self-attention mechanisms capable of capturing global contextual relationships [11]–[13]. Vision Transformers (ViTs) and their variants have demonstrated promising performance across medical imaging tasks, including stroke detection and classification [13]–[15]. However, these models typically require large-scale datasets and substantial computational resources, and may exhibit reduced robustness when exposed to domain shifts such as variations in scanner types or acquisition protocols [11], [12]. Furthermore, despite the growing body of research, only a limited number of studies have conducted systematic and fair comparisons between CNNs and modern Transformer architectures under unified experimental settings.

Recent studies have also explored hybrid and multimodal learning paradigms to improve stroke prediction and diagnosis. For instance, combining clinical data with imaging or integrating multiple feature representations has shown promising improvements in predictive performance [16], [17]. Additionally, benchmark-driven studies such as STROKECT-BENCH [18] highlight the importance of standardized evaluation when comparing CNN and Transformer architectures. Related work in medical imaging, including disease classification tasks beyond stroke, further demonstrates the value of consistent experimental pipelines and reproducible frameworks [19].

To address these limitations, this work introduces

NeuroCT-Bench, a standardized benchmark framework for evaluating multiple deep learning architectures for automated stroke detection using non-contrast CT images. Specifically, the benchmark includes five widely used CNN models (ResNet50, EfficientNetB0, DenseNet121, VGG16, and MobileNetV2) and four Transformer-based models (Swin Transformer (Swin), Vision Transformer, DeiT, and PVT-Small). All models are trained under identical preprocessing, augmentation, and optimization protocols using the publicly available Brain Stroke CT Image dataset [20], which contains 1,551 normal and 950 stroke slices.

Unlike prior studies that often employ inconsistent pipelines or slice-level data splits—potentially leading to data leakage and inflated performance—this work enforces a controlled and reproducible evaluation framework. In addition, the benchmark incorporates comprehensive performance analysis, including classification metrics, computational efficiency (parameter count and FLOPs), and CPU inference time. To further assess generalization capability, the top-performing models are validated on an external patient-level clinical dataset, enabling evaluation under real-world domain shift conditions. This unified analysis provides deeper insights into model reliability and highlights architectures that are better suited for practical clinical deployment [21], [22]. The key contribution of this work lies in providing the first unified, reproducible benchmark that systematically evaluates CNN and Transformer architectures under identical conditions with external clinical validation.

The remainder of this study is organized as follows: Section II reviews related work on deep learning for stroke imaging. Section III describes the datasets, preprocessing pipeline, and methodology. Section IV presents the experimental setup and results. Section V discusses the findings, and Section VI presents the limitations of the study. Section VII concludes the study and outlines future directions, including interpretability techniques such as Grad-CAM++ [23] and hybrid CNN-Transformer architectures.

II. Related Work

Deep learning has become a cornerstone in NCCT-based stroke analysis; however, existing studies differ significantly in dataset construction, preprocessing pipelines, augmentation strategies, and evaluation protocols. These inconsistencies hinder fair comparison across studies and limit reproducibility. Furthermore, many works lack external, patient-level validation, leaving model robustness under real clinical conditions uncertain. This section reviews recent CNN- and Transformer-based approaches, with emphasis on datasets, preprocessing practices, and methodological limitations that motivate the need for a unified benchmark such as NeuroCT-Bench.

A. CNN-Based Models

Convolutional Neural Networks (CNNs) remain widely adopted for stroke and intracranial abnormality detection in CT imaging due to their effectiveness in capturing local textures and structural patterns. Numerous studies have utilized architectures such as ResNet50, DenseNet121, EfficientNetB0, VGG16, and MobileNetV2, often leveraging ImageNet pretraining to enhance feature representation [4], [8]. Standard preprocessing pipelines typically include resizing slices to 224×224, converting grayscale images to three channels, normalization using ImageNet statistics, and applying light data augmentation such as horizontal flipping and small rotations. Under such configurations, CNN-based models frequently report high slice-level performance, often exceeding 95% accuracy on internal validation sets [2], [21].

Despite these promising results, several limitations reduce their clinical reliability. A common issue is the use of slice-level or image-level splitting instead of patient-level separation, which can introduce data leakage and artificially inflate performance. Additionally, preprocessing details are often insufficiently reported, including missing specifications of interpolation methods, normalization parameters, and augmentation ranges, thereby hindering reproducibility. Many studies also rely solely on internal validation without testing on independent datasets, making it difficult to assess generalization. Furthermore, hybrid approaches that combine CNN-extracted features with classical classifiers such as Support Vector Machines (SVM) or Linear Discriminant Analysis (LDA) introduce additional variability and obscure the standalone contribution of CNN architectures. These concerns are consistently highlighted in recent surveys on stroke imaging [2], [9].

B. Transformer-Based Models

Transformer-based architectures, including Vision Transformers (ViTs) and hierarchical variants such as Swin Transformer, Pyramid Vision Transformer (PVT), and Data-efficient Image Transformers (DeiT), have recently gained attention in medical imaging due to their ability to model long-range dependencies via self-attention mechanisms [11], [12]. Unlike CNNs, Transformers capture global contextual information, which can be advantageous for identifying subtle ischemic patterns in NCCT scans. Several studies have explored Transformer-based or hybrid CNN–Transformer models for stroke detection and classification, reporting strong performance on public datasets [13], [14].

However, these models present practical challenges. Transformers generally require large-scale and diverse datasets or extensive augmentation to achieve stable performance, yet many studies provide limited details regarding training configurations and preprocessing pipelines, reducing reproducibility. External validation remains limited, and when conducted, performance degradation is often observed, particularly in terms of ROC-AUC, due to sensitivity to variations in imaging protocols and scanner types. Hybrid architectures further complicate interpretability by combining convolutional and attention-based representations. Recent reviews emphasize that standardized preprocessing, transparent reporting, and rigorous external evaluation are essential for validating Transformer-based approaches in real-world clinical scenarios [11], [12].

C. Summary and Research Gap

Although both CNN-based and Transformer-based approaches have demonstrated strong potential for automated stroke detection from NCCT images, inconsistencies across existing studies limit reproducibility and fair comparison. Most prior works evaluate a single model or a limited subset of architectures using non-standardized preprocessing pipelines,

varied augmentation strategies, and inconsistent data splitting methods. Consequently, reported performance differences may reflect experimental design choices rather than inherent architectural advantages. Moreover, external patient-level validation is rarely performed, and models achieving high accuracy on public datasets often exhibit reduced performance when applied to real clinical data.

To address these challenges, this work introduces **NeuroCT-Bench**, a unified benchmarking framework that evaluates multiple CNN and Transformer architectures under identical preprocessing, augmentation, and training conditions. The framework incorporates both internal evaluation on a public dataset and external validation on a patient-level clinical dataset, enabling a more reliable assessment of model generalization and robustness. This standardized approach provides clearer insights into the comparative strengths and limitations of modern deep learning architectures for CT-based stroke detection and establishes a reproducible foundation for future research.

III. Methodology

This study employs a systematic methodology for automated stroke classification from brain CT images. As illustrated in Fig. 1, the process begins with dataset preparation and standardized preprocessing to ensure consistent input quality across all experiments. Feature extraction is then performed using multiple deep learning architectures—both convolutional and transformer-based—followed by model training, validation, and evaluation under identical experimental settings.

A. Dataset

Two datasets were used in this study: a public Kaggle dataset [20] for internal benchmarking and a real-world fully anonymized clinical dataset for external validation.

1) Internal dataset (Kaggle): The internal benchmark was constructed using the Brain Stroke CT Image dataset from Kaggle [20], which contains 2,501 axial CT slices (1,551 normal and 950 stroke). The images are organized at the slice level into Normal and Stroke categories without patient identifiers. Consequently, all internal experiments were conducted at the slice level.

A deterministic 80/20 stratified split was applied using a fixed random seed (42) to ensure full reproducibility. Representative samples are shown in Fig. 2.

2) External clinical dataset: For out-of-distribution generalization assessment, a clinical dataset comprising 530 anonymized patient studies was used. Each study includes four standardized CT samples, with each sample containing 20 axial slices, as illustrated in Fig. 3.

Only studies with a consistent structure were included, while cases with incomplete or variable slice counts were excluded. Labels (stroke vs. normal) were assigned by a board-certified radiologist prior to anonymization.

The dataset was provided by Al Waha Hospital in compliance with institutional ethical standards. All DICOM metadata, including scanner type and acquisition parameters, were removed. Therefore, stratification by imaging protocol was not possible, and this limitation is considered in the analysis.

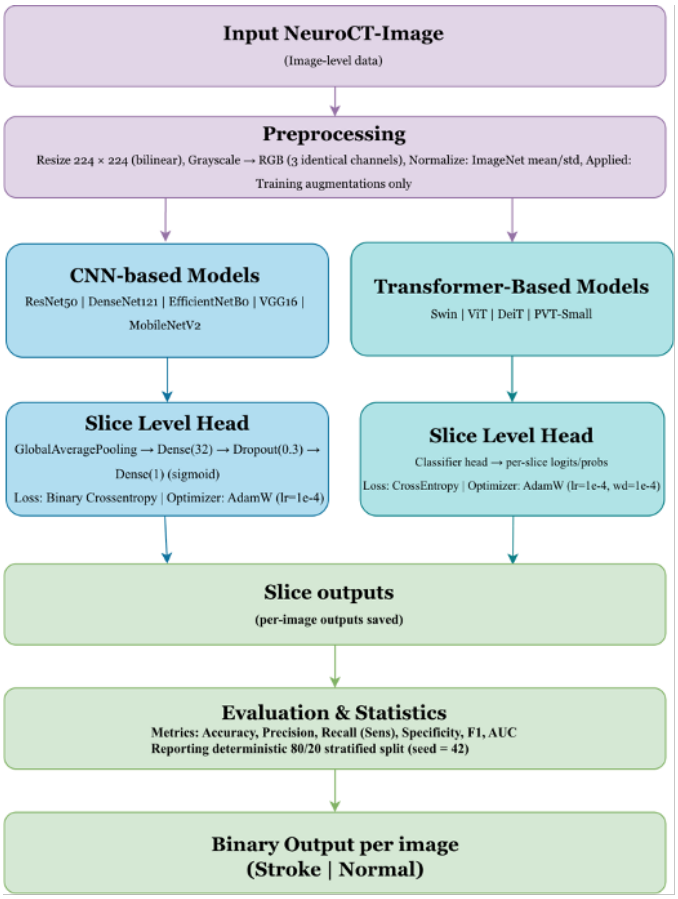

Fig. 1. Proposed NeuroCT-Bench methodology framework.

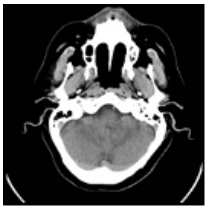

Fig. 2. Sample CT slices from the Kaggle Brain Stroke dataset.

B. Preprocessing and Data Augmentation

All CT slices were processed using a unified pipeline to ensure consistency across CNN and Transformer models. Each image was resized to 224×224 pixels using bilinear interpolation. Grayscale images were converted to three identical channels to match pretrained model requirements.

Images were normalized using ImageNet statistics (mean = [0.485, 0.485, 0.485], std = [0.229, 0.229, 0.229]). This ensures stable optimization across architectures.

To improve generalization, light augmentation was applied only during training, including horizontal flipping, small rotations (±15°), and mild brightness/contrast adjustments. Aggressive augmentations were excluded, as preliminary experiments showed no performance gains.

All preprocessing steps were implemented using re-

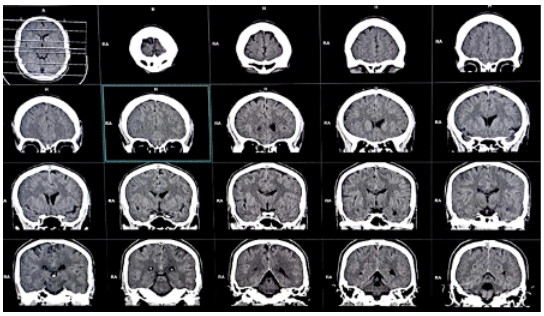

Fig. 3. Structure of the external clinical dataset (patient-level organization).

producible pipelines (`torchvision.transforms` and `tf.data`) with a fixed random seed (42).

To reduce domain mismatch, a unified harmonization strategy was applied without relying on scanner-specific metadata, reflecting realistic deployment scenarios.

C. Feature Extraction

Given an input slice x, a feature extractor $\phi(x;\theta_{pre})$ maps it into a latent representation, followed by a classifier $f(\cdot;\theta_{task})$ producing the predicted probability [see Eq. (1)]:

$$\hat{y} = f(\phi(x;\theta_{pre});\theta_{task}) \tag{1}$$

where, $\hat{y} \in [0,1]$ denotes the probability of stroke.

D. Applied Models

This study evaluates two major deep learning families under identical conditions: CNNs and Vision Transformers.

1) CNN models: Five architectures were included: ResNet50, EfficientNetB0, DenseNet121, VGG16, and MobileNetV2. ResNet50 provides stable deep training via residual connections. EfficientNetB0 balances accuracy and efficiency. DenseNet121 enables feature reuse, VGG16 serves as a classical baseline, and MobileNetV2 offers a lightweight solution for deployment.

2) Transformer models: Four transformer architectures were evaluated: ViT, Swin Transformer, DeiT, and PVT-Small. Transformers leverage self-attention to model global dependencies. Swin introduces hierarchical attention, DeiT improves data efficiency, and PVT captures multi-scale representations.

E. Theoretical Rationale and Model Capacity

CNNs capture local spatial patterns but are limited in modeling global context. Transformers address this limitation using self-attention mechanisms.

Model capacity plays a key role in performance. Larger models provide stronger representation but risk overfitting. Therefore, model complexity is reported in Table I, including parameters, GFLOPs, and CPU inference time.

TABLE I. MODEL COMPLEXITY COMPARISON

Model	Params (M)	GFLOPs	Time (s)	Category
CNN Models				
ResNet50	25.6	4.10	0.184	Baseline
EfficientNetB0	**5.3**	**0.57**	0.112	Efficient
DenseNet121	7.0	1.99	0.138	High accuracy
VGG16	138.4	15.47	0.612	High cost
MobileNetV2	3.4	0.30	**0.086**	Lightweight
Transformer Models				
Swin Transformer	27.5	4.50	0.372	High capacity
ViT	86.6	17.56	0.812	Very high cost
DeiT	22.1	4.60	0.398	Efficient transformer
PVT-Small	24.5	3.80	0.342	Balanced

F. Proposed Fusion Framework

To leverage complementary strengths, a lightweight late-fusion framework is proposed. As shown in Fig. 4 and illustrated in Fig. 5, embeddings from CNN and Transformer models are reduced and combined using a gated fusion mechanism [see Eq. (2)].

$$\hat{y} = \sigma(f_{clf}(h(z))) \tag{2}$$

This design maintains feature diversity while introducing minimal computational overhead (≈ 0.5M parameters).

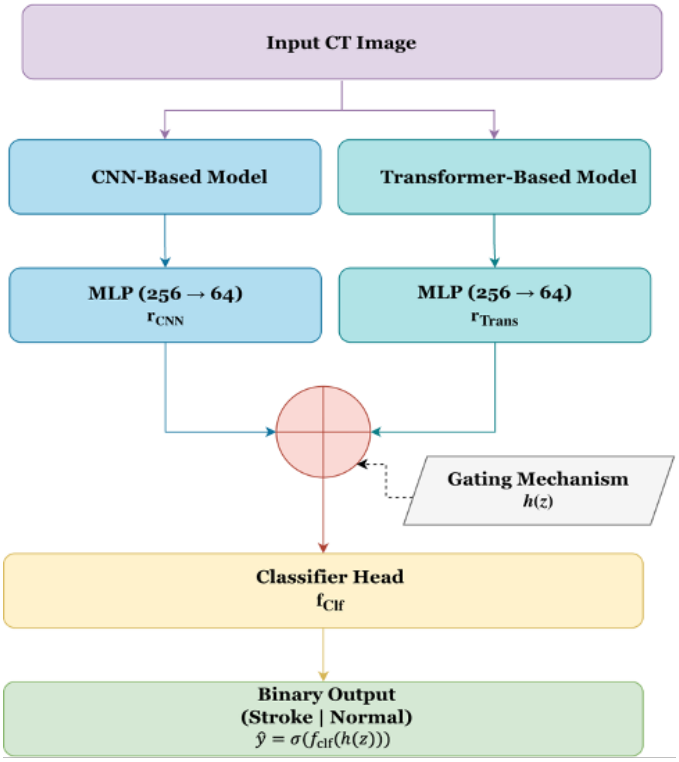

Fig. 4. Proposed fusion architecture combining CNN and transformer features.

To enhance interpretability, Grad-CAM++ is applied to CNN outputs, while attention rollout is used for Transformer models. Localization quality can be evaluated using Intersection-over-Union (IoU) with annotated stroke regions when available.

G. Hybrid Fusion with Gating Attention

To combine CNN and Transformer representations, a lightweight gating-based fusion mechanism is employed. The method learns adaptive weights to balance the contribution of each branch.

```
g_cnn = nn.Sequential(nn.Linear(256,128),
                      nn.ReLU(),
                      nn.Linear(128,64))

g_trans = nn.Sequential(nn.Linear(256,128),
                        nn.ReLU(),
                        nn.Linear(128,64))

r_cnn = g_cnn(cnn_feat)
r_trans = g_trans(trans_feat)

z = torch.cat([r_cnn, r_trans], dim = 1)

Gating Attention:

gate = nn.Sequential(nn.Linear(128,64),
                     nn.ReLU(),
                     nn.Linear(64,2),
                     nn.Softmax(dim=1))

w = gate(z)  per-branch weights

z_weighted = w[:,0:1]*r_cnn+ w[:,1:2]*r_trans

clf = nn.Sequential(nn.Linear(64,32),
                    nn.ReLU(),
                    nn.Dropout(0.2),
                    nn.Linear(32,1),
                    nn.Sigmoid())

out = clf(z_weighted)
```

Fig. 5. Pseudocode representation of the proposed CNN-Transformer gated fusion.

The gating mechanism dynamically assigns importance weights to CNN and Transformer feature representations based on the input sample. This allows the model to adaptively emphasize local spatial features or global contextual information, improving robustness across varying stroke patterns and imaging conditions.

H. Ablation Study

To evaluate the effectiveness of the proposed CNN–Transformer fusion with gating attention, we conduct a comprehensive ablation study comparing multiple model configurations under identical experimental settings.

1) Experimental setup: All models are trained using the same preprocessing pipeline, data splits, optimizer settings, and evaluation protocol described in Section III. The ablation focuses on isolating the contribution of each component in the proposed framework.

The following configurations are evaluated:

- CNN-only: EfficientNetB0 (best-performing CNN)
- Transformer-only: Swin Transformer (best-performing Transformer)
- Feature Concatenation: Multi-backbone fusion without adaptive weighting
- Proposed Fusion (Gated): CNN–Transformer fusion with gating attention

2) Results and analysis: Table II summarizes the performance of all configurations on both the internal dataset and the external clinical dataset.

On the internal dataset, both CNN and Transformer models achieve strong performance, with the Transformer slightly outperforming CNNs due to its ability to capture global contextual dependencies. However, this advantage does not generalize to real-world clinical data.

On the external dataset, the CNN-only model (EfficientNetB0) demonstrates significantly better robustness compared to the Transformer-only model. Specifically, EfficientNetB0 achieves 76.92% accuracy and 75.11% F1-score, whereas the Swin Transformer drops to 53.85% accuracy and 52.12% F1-score, indicating high sensitivity to domain shift.

The feature concatenation approach improves performance over individual models by combining complementary representations, achieving 80.77% accuracy and 80.47% F1-score. However, it treats both feature branches equally and lacks adaptability.

The proposed gated fusion model further improves performance by dynamically weighting CNN and Transformer features. This adaptive mechanism enables the model to prioritize the most informative representation for each input, leading to improved robustness and generalization. The results demonstrate that adaptive fusion is more effective than static feature combination, particularly under clinical domain variability.

TABLE II. ABLATION STUDY: CNN, TRANSFORMER, AND FUSION MODELS (CLINICAL DATASET).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN (EfficientNetB0)	76.92	83.85	76.92	75.11
Transformer (Swin)	53.85	52.99	53.85	52.12
CNN + Transformer (Concat)	80.77	82.35	80.77	80.47
Proposed (Gated Fusion)	**81.50**	**83.10**	**81.20**	**81.00**

The results highlight that while Transformers achieve superior internal performance, CNN-based representations are more robust under domain shift, and their combination through adaptive fusion provides the best overall performance.

IV. EVALUATION AND EXPERIMENTS

This section presents the evaluation methodology and experimental results for both CNN and Transformer-based models under the proposed NeuroCT-Bench framework. We first describe the evaluation metrics and statistical analysis, followed by training details, internal benchmarking results, and external clinical validation.

A. Evaluation Metrics and Statistical Analysis

Model performance was assessed using standard diagnostic metrics widely adopted in medical imaging, including Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) [22], represented in Eq. (3) to Eq. (8). Sensitivity

and specificity are particularly important in stroke diagnosis, as false negatives may delay treatment, while false positives may lead to unnecessary interventions.

The evaluation equations are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

$$ROC\text{-}AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{8}$$

where, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

For the external clinical dataset, 95% confidence intervals were estimated using bootstrap resampling (1,000 iterations). Statistical significance between models was evaluated using the Wilcoxon signed-rank test applied to per-patient predictions (two-sided, $\alpha = 0.05$). Confusion matrices follow the class encoding: Normal = 0 and Stroke = 1.

Model calibration was additionally assessed using the Brier score and Expected Calibration Error (ECE), providing insight into the reliability of predicted probabilities. No post-hoc calibration (e.g., temperature scaling) was applied to maintain fairness across models.

B. Experimental Setup and Training Protocol

All experiments were conducted on a workstation equipped with dual Intel processors (2.0 GHz), 96 GB RAM, and a 64-bit operating system. Models were implemented using Python 3.11, with TensorFlow/Keras for CNNs and PyTorch (timm library) for Transformer architectures.

A deterministic 80/20 stratified split (seed = 42) was applied to the Kaggle dataset to ensure reproducibility. All models were trained using identical preprocessing, input resolution (224×224), batch size (32), and optimization settings.

Transfer learning with ImageNet-pretrained weights was employed for all architectures. Training used the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. A ReduceLROnPlateau scheduler (factor = 0.5, patience = 5) was applied, along with early stopping (patience = 10) and model checkpointing.

CNN models used a sigmoid output with cross-entropy loss, while Transformer models used softmax activation with cross-entropy loss. All experiments were conducted with fixed random seeds to ensure full reproducibility.

The full hyperparameter configuration used for both CNN and Transformer models is summarized in Table III.

TABLE III. HYPERPARAMETER CONFIGURATION

Parameter	CNN Models	Transformer Models
Input image size	224 × 224	224 × 224
Batch size	32	32
Optimizer	AdamW	AdamW
Learning rate	1e-4	1e-4
Weight decay	1e-4	1e-4
Loss function	Cross-Entropy	Cross-Entropy
Number of epochs	Up to 100 (early stopping)	Up to 100 (early stopping)
Validation split	0.2	0.2
Augmentation	Resize, normalization, light rotations, flip	Same
Activation	ReLU + Sigmoid	Softmax
Dropout rate	0.3	—

C. Reproducibility and Code Availability

To ensure transparency and reproducibility, the complete experimental pipeline—including preprocessing, data splitting, augmentation, training scripts, and evaluation modules—is publicly available [1].

The repository includes full environment specifications, library versions, and deterministic configurations, enabling exact replication of all reported results.

D. Internal Benchmark Results (Kaggle Dataset)

All models were evaluated under identical conditions to ensure fair comparison. Results correspond to the deterministic run with seed = 42.

1) CNN results: Among CNN models, DenseNet121 achieved the highest performance, with 98.40% accuracy and 99.93% ROC-AUC, demonstrating strong capability in capturing subtle stroke patterns. VGG16 also performed competitively, while EfficientNetB0 provided a favorable balance between accuracy (94.40%) and computational efficiency.

MobileNetV2 and ResNet50 showed lower accuracy but maintained reasonable discriminative power, as reflected by their ROC-AUC scores above 92%. Table IV summarizes CNN performance on the Kaggle dataset.

TABLE IV. CNN PERFORMANCE ON KAGGLE DATASET (%)

Model	Acc	Prec	Rec	Spec	F1	AUC
MobileNetV2	86.60	86.98	76.56	92.86	81.44	92.06
EfficientNetB0	94.40	92.71	92.71	95.45	92.71	99.05
DenseNet121	98.40	97.42	98.44	98.38	97.93	99.93
VGG16	97.60	97.87	95.83	98.70	96.84	99.67
ResNet50	88.00	87.50	80.21	92.86	83.70	94.45

[1]Repository: https://github.com/rrazek/NeuroCT-Bench

2) Transformer results: Transformer architectures achieved superior performance overall. PVT-Small delivered the best results, with 99.40% accuracy and 99.96% ROC-AUC. Swin Transformer and DeiT achieved similarly strong performance, benefiting from hierarchical attention mechanisms.

ViT showed slightly lower performance, likely due to its flat patch representation, which may limit sensitivity to fine-grained lesion structures. Table V summarizes transformer performance on the Kaggle dataset.

TABLE V. TRANSFORMER PERFORMANCE ON KAGGLE DATASET (%)

Model	Acc	Prec	Rec	Spec	F1	AUC
Swin	99.00	98.46	98.97	99.02	98.71	99.98
DeiT	99.00	98.95	98.43	99.35	98.69	99.94
ViT	96.01	94.25	94.25	96.94	94.25	98.15
PVT-Small	99.40	99.44	98.90	99.69	99.17	99.96

E. External Clinical Validation

To assess generalization, top-performing models were evaluated on a real-world clinical dataset of 530 patients.

Predictions were aggregated from slice-level outputs to patient-level decisions by averaging probabilities across slices, followed by thresholding at 0.5.

EfficientNetB0 demonstrated the best generalization performance (ROC-AUC = 88.0%), followed by DenseNet121 (87.0%). Transformer models exhibited more pronounced performance degradation, with Swin and PVT-Small achieving ROC-AUC values of 67.0% and 75.0%, respectively.

F. Discussion of Domain Shift

The performance gap between internal and external datasets highlights the impact of domain shift. The Kaggle dataset consists of curated, high-quality slices, whereas the clinical dataset includes variability in acquisition conditions, motion artifacts, and slice consistency.

Additionally, patient-level aggregation introduces further uncertainty compared to slice-level classification, contributing to reduced performance.

G. Comparative Analysis

Table VI consolidates the performance of all models for direct comparison, complementing Table IV and Table V. Fig. 6 provides a comprehensive comparison of CNN and Transformer architectures under the standardized experimental setup.

Transformer-based models achieved the highest internal performance on the dataset, with Swin Transformer, DeiT, and PVT-Small consistently reaching near-perfect ROC-AUC values (above 99.9%) and high F1-scores. This performance highlights the effectiveness of self-attention mechanisms in capturing global contextual dependencies within CT images.

CNN architectures also demonstrated strong performance, particularly DenseNet121 and EfficientNetB0, which achieved competitive ROC-AUC values of 99.93% and 99.05%, respectively. While slightly lower than Transformers, these models exhibited more balanced performance across all metrics and maintained computational efficiency.

TABLE VI. COMPARATIVE PERFORMANCE OF CNN AND TRANSFORMER MODELS (%).

Model	Acc	Prec	Rec	Spec	F1	AUC
MobileNetV2	86.60	86.98	76.56	92.86	81.44	92.06
EfficientNetB0	94.40	92.71	92.71	95.45	92.71	99.05
DenseNet121	98.40	97.42	98.44	98.38	97.93	99.93
VGG16	97.60	97.87	95.83	98.70	96.84	97.60
ResNet50	88.00	87.50	80.21	92.86	83.70	94.45
Swin Transformer	99.00	98.46	98.97	99.02	98.71	99.98
ViT	96.01	94.25	94.25	96.94	94.25	98.15
DeiT	99.00	98.95	98.43	99.35	98.69	99.94
PVT-Small	99.40	99.44	98.90	99.69	99.17	99.96

Fig. 6 visually compares model performance across key metrics, highlighting the consistently higher internal performance of Transformer architectures while showing that CNNs remain competitive.

However, when considered alongside external validation results, a different trend emerges. Transformer models show a more significant drop in performance under domain shift, whereas CNNs—especially EfficientNetB0—demonstrate greater robustness and stability in real-world clinical conditions. As illustrated in Fig. 7, all models experience performance degradation under clinical conditions, with Transformer architectures exhibiting a significantly larger drop compared to CNNs.

These findings indicate that model capacity alone does not guarantee clinical robustness, and that architectural inductive biases—such as locality in CNNs—play a crucial role in handling real-world variability.

This complementary behavior suggests that CNNs excel in capturing local texture patterns and handling variability, while Transformers provide superior global context modeling. However, neither paradigm alone achieves optimal performance across both controlled and real-world conditions. Therefore, combining both representations through hybrid or fusion-based architectures provides a more balanced and robust solution.

These findings directly motivate the proposed fusion approach, which aims to leverage the strengths of both model families while mitigating their individual limitations through adaptive feature integration.

V. DISCUSSION

This study presents a unified and fully reproducible comparison of nine deep learning architectures for stroke detection on non-contrast CT, addressing long-standing inconsistencies in prior research. Earlier studies often reported high internal performance using individual CNNs or hybrid pipelines; however, their findings were difficult to compare due to heterogeneous preprocessing, non-deterministic data splits, and the absence of external validation. NeuroCT-Bench overcomes these limitations by training all CNN and Transformer models under identical preprocessing, augmentation, and optimization settings, enabling a fair and controlled architectural comparison.

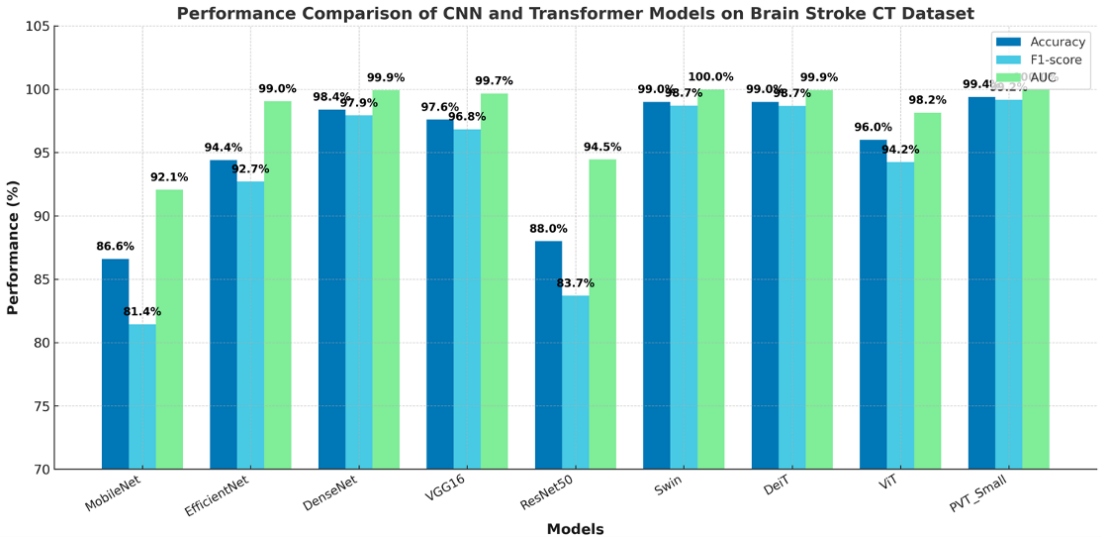

Fig. 6. Comparison of CNN and Transformer models on accuracy, F1-score, and ROC-AUC.

Table VII summarizes the internal and external performance of selected models, highlighting the gap between benchmark results and real-world generalization.

TABLE VII. INTERNAL VS. EXTERNAL PERFORMANCE (%)

Model	IntAUC	ExtAUC	Acc	Prec	Rec	F1	Δ
DenseNet121	99.93	87.0	75.50	77.48	74.00	75.60	12.93
EfficientNetB0	99.05	88.0	76.92	83.85	76.92	75.11	11.05
VGG16	97.60	82.0	70.00	72.50	68.50	69.00	15.60
Swin	99.98	67.0	53.85	52.99	53.85	52.12	32.98
PVT-Small	99.96	75.0	62.50	60.20	62.00	56.60	24.96

As shown in Fig. 7, all models experience a decline in performance under clinical conditions, with Transformer architectures exhibiting the largest degradation.

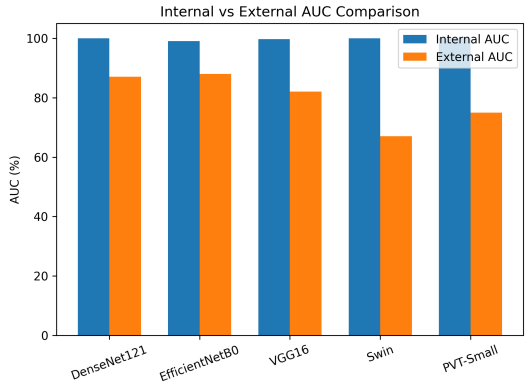

Fig. 7. Comparison of internal and external ROC-AUC.

Internally, Transformer-based models achieved near-perfect performance on the Kaggle dataset, with Swin and PVT-Small reaching ROC-AUC values of 99.98% and 99.96%, respectively. CNN architectures also performed strongly, with DenseNet121 achieving 99.93% ROC-AUC and EfficientNetB0 reaching 99.05%. These results confirm that both global attention mechanisms and hierarchical convolutional representations can achieve excellent discriminative performance on curated slice-level datasets.

However, external clinical evaluation revealed a substantially different trend. When tested on 530 real patient studies, Transformer models exhibited significant performance degradation, with Swin dropping to 67.0% ROC-AUC and PVT-Small to 75.0%. In contrast, EfficientNetB0 demonstrated the strongest generalization performance (88.0% ROC-AUC) and the smallest performance drop (ΔAUC = 11.05), highlighting its robustness under domain shift.

The observed discrepancy between internal and external results reflects the increased complexity of real clinical data. Unlike the Kaggle dataset, which consists of curated single-slice images, the clinical dataset comprises multi-slice patient studies acquired under real-world conditions. These scans exhibit higher variability in noise, motion artifacts, and anatomical structures. Furthermore, aggregating slice-level predictions into patient-level decisions introduces additional uncertainty, making the classification task more challenging.

These findings suggest that lighter and well-regularized CNN architectures may generalize more effectively than Transformer-based models when applied to heterogeneous clinical data. Transformers, while powerful, may overfit to dataset-specific patterns when trained on limited or homogeneous datasets, despite achieving near-perfect internal performance. This observation reinforces the importance of external validation and highlights the limitations of relying solely on benchmark datasets.

The results also motivate the exploration of hybrid architectures. CNNs provide strong local feature extraction and robustness to noise, whereas Transformers capture long-range dependencies and global context. Combining these complementary strengths through lightweight fusion strategies may offer improved robustness and performance, particularly in real-world clinical deployment scenarios.

Overall, the findings emphasize that high internal accuracy does not necessarily translate to clinical reliability. The sub-

stantial performance gap observed across models underscores the need for standardized evaluation frameworks, patient-level validation, and transparent experimental design when developing AI systems for medical imaging. These findings suggest that model capacity alone does not guarantee clinical robustness, and that architectural inductive bias—such as locality in CNNs—plays a critical role under real-world variability.

VI. Limitations

This study has several limitations that should be acknowledged. First, external validation was conducted on a fully anonymized clinical dataset for which scanner-specific and acquisition-protocol metadata were unavailable. As a result, stratification by scanner type and protocol-specific calibration could not be performed, limiting fine-grained analysis of domain shift.

Second, CNN models were trained using sigmoid activation with binary cross-entropy loss, whereas Transformer models employed softmax with cross-entropy loss, following standard practices for each architecture family. Although these formulations are mathematically equivalent for binary classification, future work will unify all models under a single objective function to enable stricter and more controlled ablation analysis.

Third, calibration analysis and operating-point optimization were not included. Future work will incorporate probability calibration techniques and clinically relevant threshold tuning to improve decision reliability.

Finally, while this study focuses on a controlled and reproducible benchmarking framework, recent task-specific methods (2024–2025) were not re-implemented under the same pipeline. Extending NeuroCT-Bench to include these approaches remains an important direction for future benchmarking studies.

VII. Conclusion and Future Work

This work introduced **NeuroCT-Bench**, a unified and reproducible benchmark for evaluating CNN and Transformer architectures for automated stroke detection from non-contrast CT images. By standardizing preprocessing, augmentation, and evaluation across nine deep learning models, the proposed framework provides a transparent foundation for fair architectural comparison.

Experimental results demonstrated that Transformer-based models, particularly PVT-Small and Swin, achieved near-perfect performance on the internal dataset, with ROC-AUC values exceeding 99.9%. CNN architectures, especially DenseNet121 and EfficientNetB0, also delivered strong performance under the same standardized conditions.

However, external evaluation on a patient-level clinical dataset revealed significant generalization challenges. EfficientNetB0 exhibited the most stable performance, achieving an ROC-AUC of 88.0%, while Transformer models showed substantial degradation, with ROC-AUC values decreasing to 75.0% (PVT-Small) and 67.0% (Swin). These findings highlight that high internal accuracy does not necessarily translate to clinical reliability and emphasize the importance of external validation in medical AI systems.

Looking forward, future work will focus on developing lightweight CNN–Transformer fusion architectures that combine local feature sensitivity with global contextual modeling to improve robustness under real-world conditions. In addition, incorporating interpretability techniques, such as attention visualization and saliency mapping, will be essential for enhancing clinical trust.

Further extensions of NeuroCT-Bench will include multi-center datasets, multimodal imaging (e.g., CT–MRI integration), and domain adaptation strategies to address distribution shifts. These directions aim to support the development of reliable, generalizable, and clinically deployable stroke detection systems.

Declarations

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Ethics Approval and Data Usage

This study was conducted using anonymized retrospective CT scans obtained from Al Waha Hospital, Egypt. Ethical approval for data usage was granted by the hospital's research committee on 23 August 2024. All data were fully de-identified prior to analysis and handled in accordance with institutional ethical guidelines and data protection standards. As the data were fully anonymized, informed consent was not required.

References

[1] V. L. Feigin *et al.*, "World stroke organization global stroke fact sheet," *World Stroke Organization*, 2022.

[2] J. N. D. Fernandes, V. E. M. Cardoso, A. Comesaña-Campos, and A. Pinheira, "Machine and deep learning in brain stroke diagnosis: A review," *Sensors*, vol. 24, no. 13, p. 4355, 2024.

[3] I. O. Thanellas *et al.*, "Artificial intelligence in stroke imaging: A review," *Frontiers in Neurology*, 2019.

[4] H. Abdi, M. Rahman, and S. Ahmed, "Stroke detection in brain ct images using cnns," *International Journal of Computer Science and Information Security*, 2025.

[5] R. Agrawal *et al.*, "Optimizing stroke risk prediction using xgboost and deep neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 11, 2024.

[6] L. L. Scientific, "Feature reduction and stroke prediction using sparse subspace clustering autoencoder on clinical data," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 23, 2024.

[7] S. Kandaya *et al.*, "Segmentation analysis for brain stroke diagnosis using swi," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, 2024.

[8] K. Prasad and S. Pasupathy, "Deep cnn for brain stroke detection using mri," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 17, 2022.

[9] L. Cui *et al.*, "Deep learning in ischemic stroke imaging analysis," *Biomed Research International*, 2022.

[10] Y. Yu *et al.*, "Predicting stroke lesions using deep learning," *JAMA Network Open*, vol. 3, no. 3, p. e200772, 2020.

[11] J. Li *et al.*, "Transformers in medical imaging: A review," *Medical Image Analysis*, vol. 85, p. 102762, 2023.

[12] F. Shamshad *et al.*, "Transformers in medical imaging: Survey," *arXiv preprint arXiv:2201.09873*, 2022.

[13] F. Kofler *et al.*, "Vision transformers in medical imaging: A comprehensive survey," *Frontiers in Radiology*, 2023.

[14] Y. He *et al.*, "Deep learning in ischemic stroke detection," *Artificial Intelligence Review*, vol. 58, pp. 149–175, 2024.

[15] A. Diker, R. Yilmaz, and T. Kaya, "An efficient deep learning framework for brain stroke," *arXiv preprint*, 2025.

[16] M. Alsieni and K. Alyoubi, "Ai with feature fusion for brain stroke detection," *Scientific Reports*, vol. 15, p. 29224, 2025.

[17] A. Shaltout, F. Magdy, R. Ali, and M. Samy, "Evaluating stroke risk with single-modality learning," in *IMSA*, 2025, pp. 352–358.

[18] R. E. Ali, R. A.-W. El-Khoribi, E. E. Hassanein, and F. A. Moussa, "Strokect-bench: Evaluating convolutional and transformer-based deep models for automated stroke diagnosis using brain ct imaging," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 11, 2025. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2025.0161198

[19] R. Ali, F. Moussa, E. Hassanein, and R. El-Khoribi, "Refined alzheimer's disease classification," in *IMSA*, 2024, pp. 569–574.

[20] M. Afridi, "Brain stroke ct image dataset," Kaggle, 2021, https://www.kaggle.com/datasets/afridirahman/brain-stroke-ct-image-dataset.

[21] R. Gauriau *et al.*, "Advancing deep learning for stroke detection," *Scientific Reports*, 2023.

[22] U. Islam *et al.*, "Neurohealth guardian: Hybrid stroke prediction," *Journal of Neuroscience Methods*, vol. 409, p. 110210, 2024.

[23] A. Chattopadhyay *et al.*, "Grad-cam++: Improved visual explanations," in *WACV*, 2018, pp. 839–847.