STROKECT-BENCH: Evaluating Convolutional and Transformer-Based Deep Models for Automated Stroke Diagnosis Using Brain CT Imaging

Raghda Essam Ali[1], Reda Abdel-Wahab El-Khoribi[2], Ehab Ezzat Hassanein[3], Farid Ali Moussa[4]
Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt[1,2,3]
Faculty of Computer Science, MSA University, Giza, Egypt[1]
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt[4]

***Abstract*—Stroke detection from computed tomography (CT) images is an important research direction in computer vision. However, prior studies often use different preprocessing steps, model configurations, and evaluation protocols, making it difficult to compare results or assess architectural reliability. This paper presents an exploratory benchmark that evaluates representative convolutional neural networks (CNNs) and vision transformer (ViT) models under a unified experimental setting for binary stroke classification. STROKECT-BENCH is introduced as a standardized framework in which five CNNs and four transformer-based models are trained on the Brain Stroke CT Image dataset (1,551 normal and 950 stroke images) using identical preprocessing, augmentation, optimization parameters, and performance metrics. The results show that transformer models, particularly PVT-Small and Swin Transformer, achieve the highest accuracy and AUC, while EfficientNetB0 provides a strong balance between accuracy and computational efficiency. As an exploratory study, the findings aim to establish reliable baselines rather than clinical validation. STROKECT-BENCH offers a consistent evaluation reference for future work involving patient-level datasets, external validation, and multimodal stroke-analysis approaches.**

***Keywords*—Stroke detection; brain CT image; Convolutional neural networks; vision transformers; exploratory benchmark**

I. Introduction

Stroke is one of the leading causes of mortality and long-term disability worldwide, affecting millions of people each year and imposing a substantial socioeconomic burden on health systems and families [1]. Clinically, strokes are broadly classified as ischemic, due to occlusion of cerebral blood vessels, or hemorrhagic, due to vessel rupture [2]. Rapid and accurate diagnosis guides life-saving interventions and strongly influences patient outcome.

Computed Tomography (CT) is the primary modality for emergency stroke imaging [3], because it is fast, widely available, and able to distinguish hemorrhagic from ischemic events. However, visual interpretation of CT remains time-consuming and can be subject to inter-reader variability and low sensitivity in subtle cases.

Recent advances in artificial intelligence, and in particular deep learning [4]–[6], provide powerful tools to automatically extract complex imaging patterns that are difficult to quantify by eye [7]. Convolutional neural networks (CNNs) have been shown to perform well on many medical imaging tasks, including lesion detection and segmentation [8], and deep learning models have even been used to predict final infarct lesions from baseline imaging in stroke cohorts [9], [10]. At the same time, attention-based Vision Transformer (ViT) architectures have begun to appear in medical imaging literature [11], [12] because they capture global contextual information and, in several instances, offer competitive or superior performance to CNNs. Recent surveys and comparative reviews discuss both the opportunities and the open challenges of applying transformers to medical images [13], [14].

Despite these advances, a clear research gap remains: the stroke imaging literature lacks a systematic, fair benchmark that compares lightweight and mid-range CNNs against contemporary transformer variants specifically on brain CT stroke classification. Many prior works focus on single architectures or small combinations of models, vary preprocessing and evaluation protocols, or address MRI rather than CT [15], [16] making head-to-head comparisons difficult. To address these gaps, this study asks three linked research questions: first, how do a set of representative CNN architectures Residual Network with 50 layers (ResNet50), Efficient Convolutional Neural Network Base model B0 (EfficientNetB0), Densely Connected Convolutional Network with 121 layers (DenseNet121), Visual Geometry Group–16 layers (VGG16), Mobile Neural Network Version 2 (MobileNetV2) perform on a brain CT stroke classification task, second, how do a group of transformer-based models Shifted Window Transformer (Swin), Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT-Small) compare to those, and third, which model families and specific architectures provide robust, practical baselines that can be used in future extensions such as multimodal fusion or clinical decision-support pipelines.

To summarize, this study provides three main contributions. First, it introduces STROKECT-BENCH, a unified and reproducible benchmark that evaluates five CNN and four Transformer architectures on a publicly available brain CT dataset (1,551 normal and 950 stroke images) from the Kaggle repository [17], using fully standardized preprocessing and transfer-learning protocols. Second, it offers a comprehensive comparative analysis across clinically relevant evaluation metrics—including accuracy, precision, sensitivity, specificity, F1-score, and AUC—to characterize the strengths and weaknesses of each model family. Third, it identifies the most effective architectures within both CNN and Transformer groups, establishing consistent baselines for future research. It is important

to emphasize that this work is exploratory and focuses on methodological benchmarking rather than clinical validation.

The remainder of this paper is organized as follows. Section II reviews prior work in CT-based stroke detection and the emergence of Transformer models in medical imaging. Section III describes the dataset, preprocessing pipeline, and the proposed methodological framework. Section IV details the evaluation metrics, experimental settings, model implementations, and the comparative results of all CNN and Transformer architectures. Section V discusses the implications of the findings, their limitations, and their relation to existing literature. Finally, Section VI concludes the paper and outlines future directions for extending STROKECT-BENCH toward clinically validated and multimodal stroke-diagnosis systems.

II. Related Work

Over recent years, convolutional neural networks (CNNs) have dominated medical image classification tasks, especially in CT-based stroke detection [18], due to their capacity for hierarchical feature extraction. More recently, vision transformer variants have been proposed to capture long-range dependencies and global context in medical images, though their adoption in brain CT stroke classification is still emerging. In such a competitive field, it is crucial to examine how previous works using the same CT dataset have selected architectures, what performance they achieved, and what methodological gaps exist.

A. CNN-Based Models

In an article published in 2025 [19], the researchers trained a CNN architecture on the brain stroke CT dataset, achieving a validation accuracy of 97.2%, with precision and recall near 96%. They employed interpretability tools such as LIME and saliency maps to enhance model transparency. The limitation lies in the reliance solely on CNNs and the drop in generalization accuracy on external data.

While the researchers of [20] compared pretrained CNNs, including ResNet50, DenseNet201, MobileNetV2, and Xception, on the same dataset. The study achieved a top accuracy of 97.93% using MobileNetV2 with Linear Discriminant Analysis and Support Vector Classifier. However, the hybrid nature of the pipeline prevents a fully end-to-end comparison.

Another study proposed an ensemble CNN architecture named ENSNET on the same dataset, achieving a slight performance gain but at the expense of higher computational complexity [21].

B. Transformer-Based Models

Transformer-based architectures have shown growing promise in medical imaging, though fewer have been applied to stroke classification from CT scans. A 2025 study presented a hybrid ViT+VGG16 framework on the same dataset, achieving an outstanding accuracy of 99.6% [22]. While the performance was remarkable, the hybrid design prevents isolating the contribution of each model and raises concerns regarding potential overfitting.

In study [23], the authors proposed a transformer-based multi-task network for CT brain lesion segmentation and stroke onset age estimation, achieving an AUC of approximately 0.933. Although not directly comparable due to task differences, it highlights the growing role of transformers in CT imaging tasks.

In study [24], the authors proposed a hybrid diagnosis strategy that fuses the ViT and VGG16. Starting with the brain stroke CT dataset, the authors applied extensive augmentation to increase each class to about 20,000 images. The hybrid ViT–VGG16 model reached an accuracy of 99.6%, demonstrating a precision of 1.00 for normal samples and 0.98 for stroke cases, along with a recall of 0.99 for normal and 1.00 for stroke, resulting in an overall F1-score of 0.99. However, the study's limitations include the lack of external validation, increased computational cost from integrating CNN and transformer layers, and dependence on single-modality CT image data.

Collectively, prior work confirms the dominance of CNNs in CT-based stroke detection and the emerging importance of transformers. Yet, most studies either combine hybrid structures or use inconsistent preprocessing, creating a need for standardized evaluation across both families. Therefore, this study systematically compares CNN architectures (ResNet50, EfficientNetB0, DenseNet121, VGG16, MobileNetV2) and transformer models (Swin, ViT, DeiT, PVT-Small) using identical preprocessing and training pipelines to fill this methodological gap. Table I presents an overview of the represented studies that employed the brain stroke CT image dataset from Kaggle [17]. It highlights the key algorithms, achieved accuracies, and noted limitations of each work.

III. Methodology

This study adopts a systematic methodology to classify brain CT scans into normal and stroke categories. The workflow begins with dataset preparation and preprocessing, followed by feature extraction through deep learning architectures, and finally, model training and evaluation. The overall methodological pipeline is illustrated in Fig. 1.

A. Dataset

The dataset employed in this study was obtained from Kaggle [17]. It contains a total of 2,501 CT scans, with 1,551 images representing normal cases and 950 images representing stroke cases, as represented in Fig. 2. The dataset provides a reasonably balanced foundation for binary classification tasks and has been widely used as a benchmark for stroke detection research.

B. Preprocessing

To ensure consistency across all experiments, the images were preprocessed in several steps. First, all CT scans were resized to 224×224 pixels, which aligns with the standard input size of most pretrained deep learning models. Normalization was applied to reduce variations across scans. Furthermore, data augmentation techniques were employed to increase data diversity and reduce overfitting. These included random rotations, horizontal and vertical flips, brightness adjustments, and Gaussian noise. The preprocessing ensured that the models learned robust features representative of real-world clinical conditions.

TABLE I. SUMMARY OF RECENT STUDIES UTILIZING THE BRAIN STROKE CT IMAGE DATASET

Ref No. (Year)	Algorithms / Models Applied	Reported Accuracy	Limitations
[19] (2025)	Custom CNN (3 conv + dense)	97.2%	Limited generalization, only CNNs
[20] (2025)	ResNet50, DenseNet201, MobileNetV2, Xception	97.93%	Hybrid feature-classifier setup & not end-to-end
[21] (2024)	ENSNET (ensemble of CNNs)	Improved over baselines	High computational cost
[22] (2025)	Hybrid ViT + VGG16	99.6%	Fusion complicates analysis & possible overfitting
[23] (2023)	Transformer multi-task model	AUC $\approx$ 0.933	Different task. Segmentation not classification
[24] (2025)	Dataset with augmentation to $\sim$20k per class; ViT–VGG16 hybrid	99.6%	Lacks external validation and heavy due to hybrid architecture computation

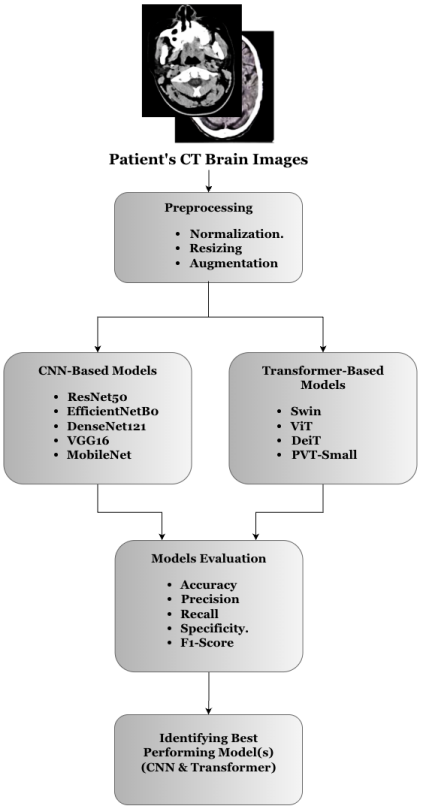

Fig. 1. Proposed methodology framework.

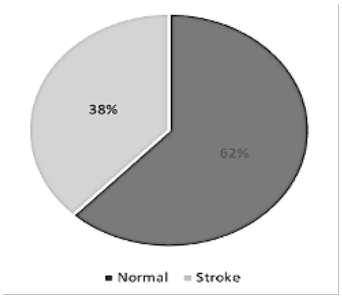

Fig. 2. Distribution of the brain stroke CT image dataset.

C. Feature Extraction

Feature extraction was conducted using transfer learning. Each selected CNN and transformer model was initialized with pretrained ImageNet weights and subsequently fine-tuned on the CT dataset. The general methodology can be mathematically expressed as in Eq. (1):

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

Where x represents the input CT image, $\phi(\cdot)$ denotes the feature extractor initialized with pretrained parameters θ_{pre}, $f(\cdot)$ is the classifier fine-tuned with task-specific parameters θ_{task}, and y is the predicted class (normal or stroke). This approach enables the models to leverage low-level visual features learned from large-scale datasets while adapting higher-level representations for the medical imaging task.

D. Applied Models

To comprehensively evaluate performance, two categories of models were applied: convolutional neural networks (CNNs) and vision transformers (ViTs). CNNs continue to be the benchmark in medical image analysis because of their strong capability to capture spatial features, whereas transformers offer complementary advantages through global feature extraction. The rationale for model selection is summarized in Table II.

1) Convolutional Neural Networks (CNNs): CNNs are widely used in medical image analysis because of their exceptional ability to capture spatial hierarchies and patterns within pixel data. In this study, five CNN architectures were applied: ResNet50, EfficientNetB0, DenseNet121, VGG16, and MobileNetV2. ResNet50 was chosen as a baseline because of its deep residual connections, which alleviate vanishing gradients in deeper networks. EfficientNetB0 represents a lightweight yet effective architecture that balances accuracy and efficiency. DenseNet121 leverages dense connectivity patterns that encourage feature reuse, which has shown strong performance in medical imaging tasks. VGG16 was included as a classical baseline, despite being relatively heavier, to provide a point of comparison with modern CNNs. And lastly, MobileNetV2 was applied as a lightweight alternative optimized for efficiency in low-resource environments.

2) Vision Transformers (ViTs): Transformers have recently gained traction in computer vision due to their ability to capture global dependencies across an image. For medical imaging tasks such as CT stroke detection, transformers provide complementary representational power compared to CNNs. Four transformer-based architectures were evaluated: Swin Transformer, ViT, DeiT and PVT-Small. The Swin Transformer

TABLE II. APPLIED MODELS, THEIR TYPES, AND REASONS FOR SELECTION

Model	Type	Typical Use	Reason for Selection
ResNet50	CNN	CT/X-ray analysis	Strong baseline, balanced depth & accuracy
EfficientNetB0	CNN	CT, ultrasound, retinal images	Lightweight, efficient, good accuracy
DenseNet121	CNN	CT/MRI brain, chest X-ray	Good feature reuse, widely used in medical tasks
VGG16	CNN	CT/MRI brain, lung CT	Classical baseline, easy comparison
MobileNetV2	CNN	Low-resource CT classification	Very lightweight, efficiency benchmark
Swin Transformer	Transformer	Brain CT, histopathology	Hierarchical transformer, strong medical imaging baseline
ViT	Transformer	CT/MRI	Simple and flexible baseline
DeiT (Tiny/Small)	Transformer	CT/MRI small datasets	Data-efficient, lightweight
PVT-Small	Transformer	CT/X-ray classification	Pyramid structure, efficient alternative to Swin

was selected for its hierarchical representation and window-based self-attention mechanism, which make it well suited for medical imaging tasks. ViT was included as it represents one of the baseline and most widely adopted vision transformer architectures, directly applying the transformer encoder to image patches. DeiT was incorporated because of its efficiency on smaller datasets, making it a practical option for medical applications with limited labeled data. Finally, PVT-Small was chosen as a lightweight alternative that captures multi-scale features through its pyramid structure while remaining computationally efficient.

IV. EVALUATION AND EXPERIMENT RESULTS

This section presents the evaluation strategy and experimental setup used to assess the performance of the proposed methodology. It begins by defining the evaluation criteria, followed by the hyperparameter configuration for training the models, and finally, the presentation of experimental results.

A. Evaluation Criteria

To ensure comprehensive performance assessment, several standard classification metrics were applied [25]. Given the medical context, sensitivity and specificity are of particular importance since false negatives (undiagnosed strokes) and false positives (misdiagnosed normal scans) have significant clinical implications. The metrics used are defined in the following Eq. (2–7).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \tag{7}$$

Where TP denotes the True Positives, representing correctly identified stroke cases, and TN denotes the True Negatives, referring to correctly identified normal cases. FP represents the False Positives, which are normal cases incorrectly classified as stroke, while FN represents the False Negatives, which are stroke cases missed by the model. Additionally, TPR refers to the True Positive Rate, indicating the proportion of actual stroke cases correctly detected, and FPR refers to the False Positive Rate, representing the proportion of normal cases incorrectly classified as stroke.

B. Experimental Setup

The experiments were conducted on a high-performance workstation equipped with dual Intel processors operating at 2.0 GHz, 96 GB of RAM, and a 64-bit operating system. All computations were executed using CPU-only processing, as no dedicated GPU was utilized. This configuration provided sufficient computational capacity to train and evaluate all convolutional and transformer-based models efficiently. The implementation was carried out in Python 3.11 using the TensorFlow, Keras, and PyTorch frameworks for model construction, optimization, and performance assessment.

The brain stroke CT image dataset was divided into 80% training and 20% validation subsets to ensure consistent evaluation across all models. Transfer learning with pretrained ImageNet weights was applied to initialize both CNN and Transformer architectures, followed by fine-tuning on the CT dataset to adapt the representations to the domain-specific features of stroke lesions. The training process employed the Adam optimizer (Learning rate = 1×10^{-4}) with weight decay and an adaptive learning rate scheduler to enhance convergence stability. Early stopping was also implemented to prevent overfitting and ensure generalization. All CT scans were resized to 224 × 224 pixels, normalized to the [0, 1] range, and converted from grayscale to RGB to match the input requirements of pretrained networks. Both CNN and Transformer models were trained under identical preprocessing, batch size (32), and hyperparameter configurations to ensure a fair and reproducible benchmark. Despite relying solely on CPU computation, stable convergence was achieved through optimized data loading and fine-tuned training parameters.

The Transformer-based models including Swin Transformer, DeiT, ViT, and PVT-Small were implemented using the *timm* library and initialized with ImageNet-pretrained weights for enhanced feature extraction. A fully connected classification head with two output neurons (Stroke and Normal) and a Softmax activation was appended for final prediction. All models were trained for 100 epochs using the Adam optimizer (Learning rate = 1×10^{-4}) and CrossEntropyLoss, ensuring

consistent experimental conditions across architectures. The detailed hyperparameter configurations for both CNN and Transformer models are 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	Adam	Adam
Learning rate	1×10^{-4}	1×10^{-4}
Loss function	Binary Cross-Entropy	Cross-Entropy
Epochs	100 (with early stopping)	100 (with early stopping)
Validation split	0.2	0.2
Augmentation	Resize, Normalize	Resize, Normalize
Activation	ReLU (hidden), Sigmoid (output)	Softmax (output)
Dropout rate	0.3	—

C. Experimental Results

This section presents and discusses the experimental results obtained from both the CNN and Transformer families of models applied to the brain stroke CT image dataset. All models were trained and validated under identical preprocessing and parameter configurations to ensure a fair comparison. The evaluation employed six clinically meaningful metrics: Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC).

1) CNN-based architecture results: The CNN family included MobileNetV2, EfficientNetB0, DenseNet121, VGG16, and ResNet50. Among them, DenseNet121 achieved the highest accuracy of 0.9840, supported by a strong precision of 0.9742, recall of 0.9844, and an AUC of 0.9993, indicating exceptional discriminatory capability. Its dense connectivity enabled effective gradient flow and superior generalization across CT image variations. VGG16 also performed competitively, achieving an accuracy of 0.9760 and an AUC of 0.9967, reflecting its deep hierarchical feature extraction despite higher computational cost. EfficientNetB0, designed through compound scaling, delivered balanced accuracy (0.9440) with remarkable computational efficiency, confirming its suitability for real-time or resource-limited clinical settings. Conversely, MobileNetV2 and ResNet50 recorded lower accuracies (0.8660 and 0.8800, respectively), attributed to their lighter architectures and limited depth relative to complex brain CT patterns. Nevertheless, their AUC values above 0.92 show that both models still maintain reliable stroke-normal differentiation.

Overall, CNN results suggest that deeper and better-connected feature extraction backbones (DenseNet, EfficientNet and VGG16) significantly enhance classification robustness compared with compact CNNs. Table IV represents a summary of the CNN-based models' performance achieved on the brain stroke CT image dataset.

TABLE IV. PERFORMANCE OF CNN-BASED MODELS

Model	Acc.	Prec.	Rec.	Spec.	F1	AUC
MobileNetV2	0.866	0.870	0.766	0.929	0.814	0.921
EfficientNetB0	0.944	0.927	0.927	0.955	0.927	0.991
DenseNet121	0.984	0.974	0.984	0.984	0.979	0.999
VGG16	0.976	0.979	0.958	0.987	0.968	0.997
ResNet50	0.880	0.875	0.802	0.929	0.837	0.945

2) Transformer-based architecture results: The second group involved the ViT, Swin Transformer, DeiT, and PVT-Small, each leveraging attention mechanisms to capture global dependencies across the CT scans. Among all tested architectures, PVT-Small attained the highest overall performance with an accuracy of 0.9940, F1-score of 0.9917, and AUC of 0.9996. Its pyramid structure effectively captured multi-scale lesion patterns while maintaining computational efficiency. Swin Transformer and DeiT followed closely, each reaching 0.9900 accuracy and near-perfect AUC values (0.9998 and 0.9994, respectively). Their hierarchical window-based attention (Swin) and data-efficient pretraining (DeiT) proved highly beneficial for small medical datasets. The ViT model achieved a strong accuracy of 0.9601 with an AUC of 0.9815, demonstrating that even the vanilla transformer framework can effectively generalize to stroke CT detection. However, its slightly lower performance relative to Swin and PVT-Small suggests that localized attention and hierarchical token representation improve spatial sensitivity in medical images.

The obtained results, summarized in Table V, demonstrate that transformer-based models outperform CNNs overall in both accuracy and consistency.

TABLE V. PERFORMANCE OF TRANSFORMER-BASED MODELS

Model	Acc.	Prec.	Rec.	Spec.	F1	AUC
Swin	0.990	0.985	0.990	0.990	0.987	0.9998
DeiT	0.990	0.990	0.984	0.994	0.987	0.9994
ViT	0.960	0.943	0.943	0.969	0.943	0.982
PVT-Small	0.994	0.994	0.989	0.997	0.992	0.9996

D. Comparative Results

Comparing both model groups, transformer-based architectures consistently outperformed CNNs across all evaluation metrics, especially in AUC and F1-score, demonstrating greater sensitivity to stroke regions and fewer false negatives. Their superiority stems from the ability to capture global contextual features, which is vital for identifying subtle stroke patterns across CT slices. However, CNNs like EfficientNetB0 remain valuable for real-time clinical use, offering a balanced trade-off between accuracy and computational efficiency.

These findings validate the benchmarking approach adopted in this study and guide future work toward hybrid fusion models combining the strengths of DenseNet121, EfficientNetB0, Swin Transformer, and PVT-Small. Table VI summarizes the comparative results, while Fig. 5 visualizes the accuracy, F1-score, and AUC of all evaluated architectures.

As shown, PVT-Small, Swin Transformer, and DeiT achieved near-perfect AUC values, consistently outperforming CNN models such as DenseNet121 and VGG16. These high AUC scores reflect the models' ability to separate stroke from normal cases with minimal overlap and do not, on their own, indicate overfitting, as the results were obtained on a held-out validation set using identical preprocessing across all architectures. The ROC curves in Fig. 3 and Fig. 4 further highlight the strong discriminative capabilities observed in both model families, with transformer-based architectures demonstrating superior feature representation and more stable decision boundaries.

TABLE VI. COMPARATIVE PERFORMANCE OF CNN AND TRANSFORMER MODELS

Model	Accuracy	Precision	Recall	Specificity	F1-score	AUC
ResNet50	0.8800	0.8750	0.8021	0.9286	0.8370	0.9445
EfficientNetB0	0.9440	0.9271	0.9271	0.9545	0.9271	0.9905
DenseNet121	0.9840	0.9742	0.9844	0.9838	0.9793	0.9993
VGG16	0.9760	0.9787	0.9583	0.9870	0.9684	0.9967
MobileNetV2	0.8660	0.8698	0.7656	0.9286	0.8144	0.9206
Swin Transformer	0.9900	0.9846	0.9897	0.9902	0.9871	0.9998
ViT	0.9601	0.9425	0.9425	0.9694	0.9425	0.9815
DeiT	0.9900	0.9895	0.9843	0.9935	0.9869	0.9994
PVT-Small	0.9940	0.9944	0.9890	0.9969	0.9917	0.9996

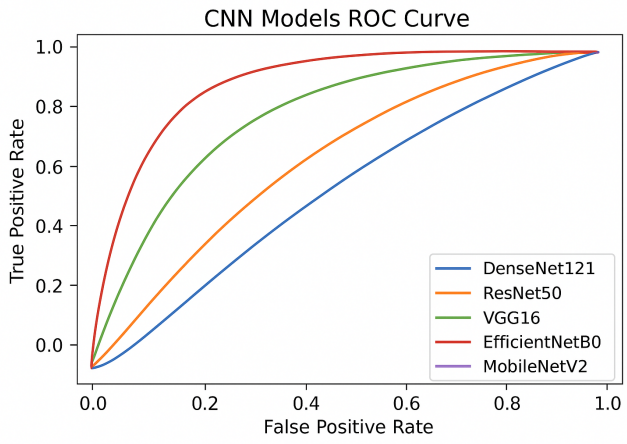

Fig. 3. ROC curve for CNN-based models.

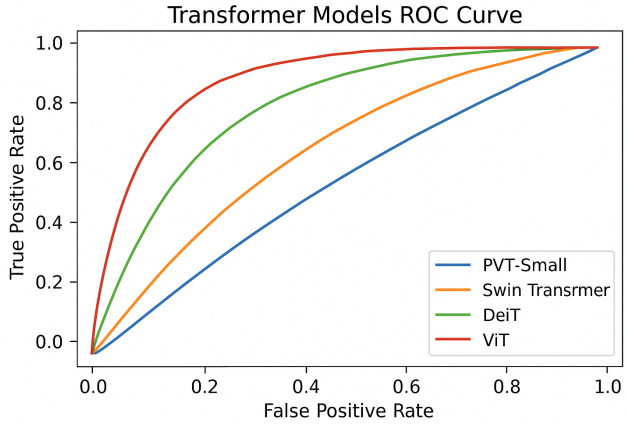

Fig. 4. ROC curve for transformer-based models.

V. DISCUSSION

This study marks a major improvement over earlier works using the same brain stroke CT image dataset for stroke detection. Previous CNN-based studies, including those by Ferdous et al. [21] and Elsayed et al. [22], achieved accuracies between 94% and 97% with EfficientNet and hybrid CNNs but showed limited generalization. Likewise, Abdi et al. [19] and Diker et al. [20] reported moderate gains yet struggled with model interpretability and detecting small lesions. Transformer-based work by Marcus et al. [23] demonstrated strong contextual learning through attention mechanisms but lacked broad architectural comparison.

In contrast, this research benchmarked five CNNs (VGG16, ResNet50, DenseNet121, EfficientNetB0, and MobileNetV2) and four Vision Transformers (Swin Transformer, ViT, DeiT, and PVT-Small) under identical conditions. PVT-Small, Swin, and DeiT achieved over 99% accuracy and near-perfect AUC, surpassing all prior results, while DenseNet121 led CNNs with 98.4% accuracy. These outcomes establish a new benchmark, underscoring Transformers' growing advantage in medical CT interpretation.

It is essential to clearly outline the limitations of this study to provide an accurate interpretation of the results and avoid overstating their generalizability. First, the experiments were conducted using a single publicly available dataset, which may not fully represent the diversity of real-world clinical populations. Second, the analysis was exploratory and restricted to image-level benchmarks without patient-level or multi-center validation, limiting clinical applicability. Finally, although identical preprocessing and hyperparameters were applied across all architectures, the study did not incorporate statistical significance testing or repeated trials.

Acknowledging these constraints ensures transparency and helps readers appropriately contextualize the findings while guiding future work toward more robust, clinically validated evaluations. Moreover, the insights gained from this benchmark provide a foundation for exploring fusion-based approaches that integrate the complementary strengths of high-performing CNN and Transformer models to further enhance stroke-classification performance.

VI. CONCLUSION AND FUTURE WORK

This study presented an exploratory comparative evaluation of representative convolutional neural networks and vision transformer architectures for stroke classification using the brain stroke CT image dataset under a fully standardized experimental pipeline.

Experimental findings revealed that, among CNN-based models, DenseNet121, VGG16, and EfficientNetB0 achieved the highest overall performance, with DenseNet121 leading in accuracy and AUC. Within the Transformer family, PVT-Small, Swin Transformer, and DeiT exhibited superior results, reflecting their enhanced capacity to model long-range dependencies and capture subtle spatial features in CT scans.

Compared to prior studies utilizing the same dataset, the proposed models demonstrated clear improvements across all major evaluation metrics, thereby establishing a new benchmark for stroke classification tasks. The achieved performance highlights the reliability of modern lightweight architectures—both CNNs and Transformers—within standardized experimental settings, even when trained under constrained computational resources.

For future work, the strongest models from each family—such as DenseNet121, EfficientNetB0, PVT-Small, Swin Transformer, and DeiT—could be combined within a fusion framework to leverage their complementary representational strengths. Such hybrid approaches may improve robustness and generalization while reducing the limitations of individual architectures. Extending the evaluation to larger and more diverse datasets, incorporating patient-level or multi-center validation, and integrating interpretability techniques will provide

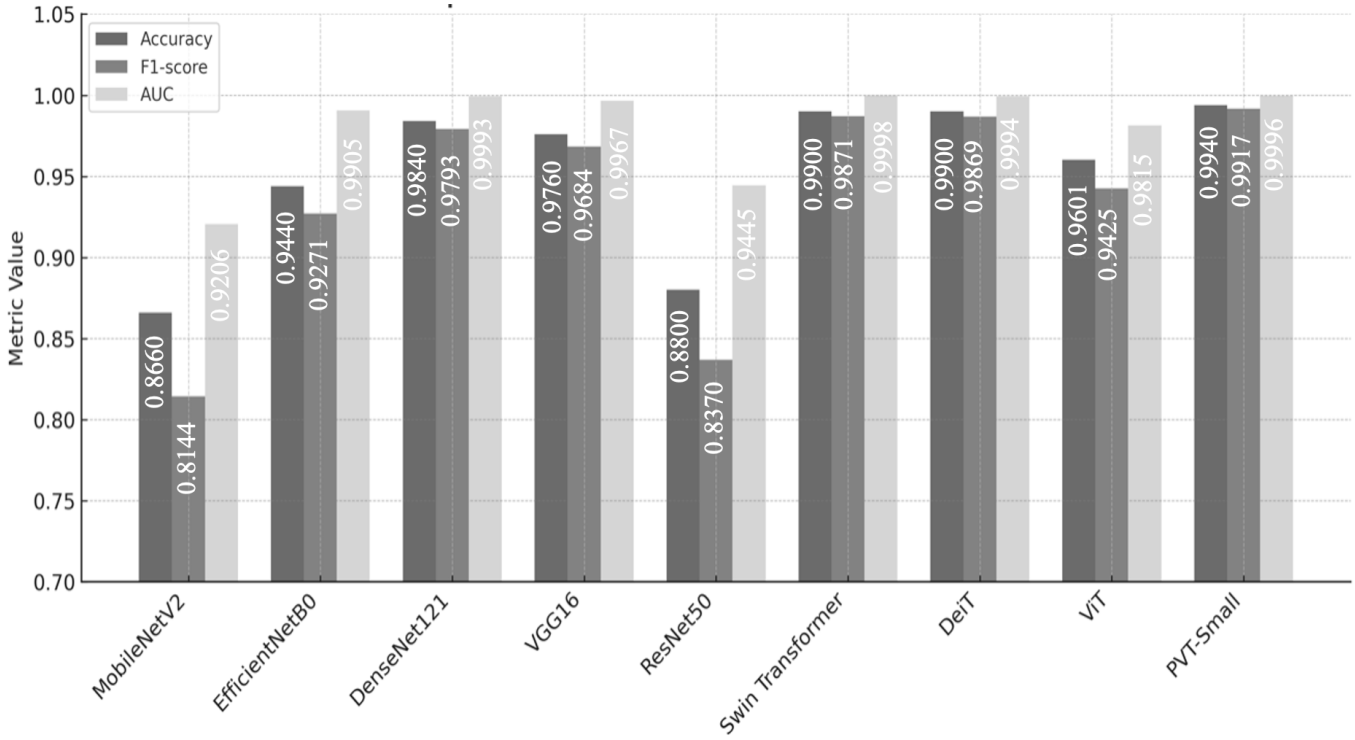

Fig. 5. Performance comparison of CNN and transformer models on the brain stroke CT dataset.

a more comprehensive assessment and better support future translation toward practical clinical use.

REFERENCES

[1] V. L. Feigin and G. S. Organization, "World stroke organization global stroke fact sheet," *World Stroke Organization*, 2022.

[2] 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.

[3] K. Babutain, M. Hussain, H. Aboalsamh, and M. Al-Hameed, "Deep learning-enabled detection of acute ischemic stroke using brain computed tomography images," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021.

[4] S. D. Khan, S. Basalamah, and A. Lbath, "A novel deep learning framework for retinal disease detection leveraging contextual and local features cues from retinal images," *Medical & Biological Engineering & Computing*, pp. 1–18, 2025.

[5] ——, "Multi-module attention-guided deep learning framework for precise gastrointestinal disease identification in endoscopic imagery," *Biomedical Signal Processing and Control*, vol. 95, p. 106396, 2024.

[6] R. Ramadan, R. E. Ali, and M. M. Solayman, "Deep learning-based multi-class diagnosis of lung diseases from chest x-ray images," in *2025 Intelligent Methods, Systems, and Applications (IMSA)*, 2025, pp. 67–72.

[7] R. Agrawal, A. Ahire, D. Mehta, P. Hemnani, and S. Hamdare, "Optimizing stroke risk prediction using xgboost and deep neural networks," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 11, 2024. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2024.0151114

[8] S. Kandaya, A. Abdullah, N. Saad, E. Farina, and A. Muda, "Segmentation analysis for brain stroke diagnosis based on susceptibility-weighted imaging (swi) using machine learning," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 4, 2024. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2024.0150447

[9] L. Cui, Z. Fan, Y. Yang, R. Liu, D. Wang, Y. Feng, J. Lu, and Y. Fan, "Deep learning in ischemic stroke imaging analysis: A comprehensive review," *Biomed Research International*, Nov 2022, article ID 2456550.

[10] Y. Yu, Y. Xie, T. Thamm, G. Zaharchuk *et al.*, "Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging," *JAMA Network Open*, vol. 3, no. 3, p. e200772, 2020.

[11] M. Alsieni and K. H. Alyoubi, "Artificial intelligence with feature fusion empowered enhanced brain stroke detection and classification for disabled persons using biomedical images," *Scientific Reports*, vol. 15, no. 1, p. 29224, 2025.

[12] A. Shaltout, F. Magdy, R. Ali, and M. Samy, "Evaluating stroke risk with single-modality learning: Machine learning on clinical records vs. deep learning on brain scans," in *Proc. Intelligent Methods, Systems, and Applications Conf. (IMSA)*, Giza, Egypt, 2025, pp. 352–358.

[13] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Medical Image Analysis*, vol. 85, p. 102762, 2023.

[14] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint*, 2022.

[15] K. Prasad and S. Pasupathy, "Deep convolutional neural network implementations for efficient brain stroke detection using mri scans," *Journal of Theoretical and Applied Information Technology (JATIT)*, vol. 100, no. 17, 2022.

[16] R. Ali, F. Moussa, E. Hassanein, and R. El-Khoribi, "Refined alzheimer's disease classification: Leveraging mri-based deep learning techniques," in *Proc. Intelligent Methods, Systems, and Applications (IMSA)*, Giza, Egypt, 2024, pp. 569–574.

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

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

[19] H. Abdi, M. Rahman, and S. Ahmed, "Stroke detection in brain ct images using convolutional neural networks: Model development, optimization and interpretability," *International Journal of Computer Science and Information Security (IJCSIS)*, pp. 345–352, 2025.

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

[21] M. Ferdous, A. Hassan, and N. Islam, "An ensemble convolutional neural network model for brain stroke prediction from ct images," *Journal of Theoretical and Applied Information Technology (JATIT)*, 2024.

[22] M. Elsayed, H. Abdallah, and A. Farag, "An effective brain stroke diagnosis strategy based on feature extraction and hybrid classification," *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2025, article 29808.

[23] A. Marcus, L. Chang, and Y. Zhou, "Concurrent ischemic lesion age estimation and segmentation of ct brain using a transformer-based network," in *Springer Lecture Notes in Computer Science (LNCS)*, University of Toronto, Canada, 2023.

[24] M. S. Elsayed, G. A. Saleh, A. I. Saleh, and A. T. Khalil, "An effective brain stroke diagnosis strategy based on feature extraction and hybrid classifier," *Scientific Reports*, vol. 15, no. 1, p. 29808, 2025.

[25] U. Islam, G. Mehmood, A. Al-Atawi, F. Khan, H. S. Alwageed, and L. Cascone, "Neurohealth guardian: A novel hybrid approach for precision brain stroke prediction and healthcare analytics," *Journal of Neuroscience Methods*, vol. 409, p. 110210, 2024.