



StrokeFuse-AttnNet: a hybrid feature fusion and self-attention model for stroke detection using neuroimages

Muhammad Asim Saleem¹ · Ashir Javeed² · Wasan Akarathanawat^{3,4} · Aurauma Chutinet^{3,4} · Nijasri Charnnarong Suwanwela^{3,4} · Pasu Kaewplung¹ · Surachai Chaitusaney¹ · Watit Benjapolakul¹

Received: 23 December 2025 / Accepted: 4 March 2026
© The Author(s) 2026

Abstract

Stroke detection and classification from computed tomography (CT) remains a critical and challenging task in medical imaging due to the complexity of lesion patterns, noise variations and unbalanced datasets. In this study, we propose a novel hybrid deep learning model, StrokeFuse-AttnNet, which integrates both global (ResNet50) and local (DenseNet121) convolutional feature extractors with a self-attention mechanism to improve spatial focus and semantic interpretability. A hierarchical feature fusion strategy concatenates multi-scale features, which are then processed by a self-attention module to highlight key stroke regions and reduce irrelevant activations. We use data augmentation and SMOTE on training samples to address imbalance and improve generalization. The proposed model was evaluated on both publicly and privately available brain CT datasets. StrokeFuse-AttnNet achieved an accuracy of 98.27% and an AUC of 0.983 on the public dataset and an accuracy of 96.04% and an AUC of 0.9501 on the private dataset. The results show that the model has higher accuracy, reliability and generalization than existing and baseline methods. The proposed model is lightweight, with only 32 million parameters and can be used in real-time clinical diagnostic processing systems that require 40 GFLOPs. The model has the potential to support radiologists in the efficient and rapid diagnosis of strokes on non-contrast CT images.

Keywords CT-based stroke detection · Hybrid feature fusion · Self-attention mechanism · Medical imaging · Deep learning

Muhammad Asim Saleem, Ashir Javeed, Wasan Akarathanawat, Aurauma Chutinet, Nijasri Charnnarong Suwanwela, Pasu Kaewplung, Surachai Chaitusaney, Watit Benjapolakul contributed equally to this work.

-
- ✉ Wasan Akarathanawat
wasan.ak@chula.ac.th
 - ✉ Pasu Kaewplung
Pasu.K@chula.ac.th
 - ✉ Watit Benjapolakul
Watit.B@chula.ac.th

¹ Department of Electrical Engineering, Faculty of Engineering, Center of Excellence in Artificial Intelligence, Machine Learning and Smart Grid Technology, Chulalongkorn University, Bangkok 10330, Thailand

² Department of Computer Science, Blekinge Institute of Technology, Blekinge, Sweden

³ Division of Neurology, Department of Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand

Introduction

Stroke is a cardiovascular disease caused by an interruption of blood supply to the brain and remains a significant health burden worldwide. It also contributes to high global rates of mortality and disability over time [1, 2]. Early and accurate diagnosis of stroke is essential to minimize its life-changing effects. Recently, artificial intelligence (AI) systems and machine learning (ML) techniques have enabled the development of new methods for assessing TIAs and strokes, particularly using CT. CT imaging is crucial for the early detection of ischaemic and haemorrhagic strokes, which require different treatment approaches [3, 4].

Automated analysis of medical imaging, particularly for stroke diagnosis, has advanced rapidly with recent developments in artificial intelligence [5]. Traditional machine learning methods, such as support vector machines and deci-

⁴ Chulalongkorn Stroke Center, Chula Neuroscience Center, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok 10330, Thailand

sion trees, rely on manually engineered high-level features derived from predefined brain regions or clinical variables. In contrast, deep learning approaches automatically learn hierarchical feature representations directly from raw neuroimages. Deep learning methods, especially convolutional neural networks (CNNs) [6], automatically learn hierarchical representations in raw neuroimages without manual feature generation, significantly reducing the number of manually created features [7]. Existing stroke diagnostic models based on deep learning have adopted architectures such as U-Net and EfficientNet for segmentation and detection and have made significant progress in identifying cerebral haemorrhage and ischaemic regions [8]. These models outperform conventional methods by detecting smaller pathological patterns that are undetectable through clinical observation. Furthermore, CNNs and texture-based deep feature extractors have significantly improved classification accuracy, demonstrating that deep learning-based systems can assist in clinical decision support [9, 10]. The proposed StrokeFuse-AttnNet model extends this deep learning-based approach by fusing multiscale features from ResNet50 and DenseNet121, which are further refined using a self-attention mechanism to enhance the reliability and interpretability of the classification. The Fig. 1 shows CT scans from the datasets, displaying both normal and stroke cases.

This study proposes a deep learning approach that integrates ResNet50, DenseNet121 and Self-Attention. The model combines these architectures to effectively process brain CT images in greyscale. The study uses both public data and private data from the Chulalongkorn Stroke Center, making it more practical. Incorporating data from the Chulalongkorn Stroke Center, King Chulalongkorn Memorial Hospital, enables the model to learn generalizable characteristics that are less represented in public datasets. The use of two datasets enhances the model's generalization, enabling it to better adapt to distinct clinical conditions. The current model builds on earlier hybrid work in feature extraction, mechanism modelling and machine learning [11]. CNN backbones with attention mechanisms are effective in addressing threats to conventional diagnostic workflows. By employing the self-attention mechanism, the accuracy and reliability of the model were improved for stroke-related abnormality detection and classification.

Deep learning plays a significant role in the analysis of CT non-contrast images, which is important for stroke diagnosis. Research using CNN architectures with robust preprocessing and hyperparameter tuning has achieved accuracies above 97% in classifying normal, ischaemic and haemorrhagic strokes [12, 13]. Recent models such as EfficientNet-B0 and ResNet50 provide further evidence supporting the clinical use of these methods. In addition to imaging, machine learning techniques applied to electronic health records (EHRs) have identified important risk factors for stroke prediction.

Stroke prediction has achieved an accuracy of 96.5% using genetic algorithms and bidirectional long short-term memory (BiLSTM) models. These approaches allow clinicians to gain insight into results while significantly reducing manual analysis [14].

Recent advances in machine learning have shown that complex, nonlinear relationships in high-dimensional biomedical data are most effectively modelled using hybrid and ensemble-based strategies rather than single-model architectures. Studies indicate that ensemble learning, attention-guided representations and multi-backbone fusion can significantly improve robustness, generalization and interpretability when handling heterogeneous clinical data [15, 16]. In particular, attention mechanisms have been successfully used to selectively emphasise informative feature interactions in nonlinear systems, while ensemble and hybrid models reduce overfitting and sensitivity to data variability [17, 18].

More recently, hybrid machine learning frameworks that combine multiple feature extractors and adaptive weighting schemes have been shown to outperform conventional deep models across a wide range of applied domains by capturing complementary structural and semantic information [19]. These findings highlight a broader trend in modern machine learning towards structured model integration and interpretable attention mechanisms precisely the challenges encountered in neuro imaging-based stroke diagnosis, where lesion appearance, anatomical variability and imaging artefacts introduce strong nonlinearity and heterogeneity.

Despite these improvements, several challenges remain, particularly in developing effective AI models suitable for resource-constrained environments. To address this, this study proposes a scalable framework that mitigates computational limitations through efficient architectural design and adaptive training strategies. This approach accommodates diversity in hospital and healthcare structures while ensuring maximum efficiency and accuracy. The model designed for stroke detection will enhance the accessibility and reliability of stroke detection solutions within clinical settings.

The main contributions of this study are as follows:

- (i) Proposes StrokeFuse-AttnNet, a hybrid framework combining ResNet50 and DenseNet121, that extracts both global and local features from grey-scale CT scan images.
- (ii) This study analyzes a hybrid feature fusion strategy using ResNet50 and DenseNet121 for more reliable visualization.
- (iii) To strengthen the fused features and identify the hidden stroke features, a self-attention mechanism is employed.
- (iv) With only 32 million parameters and a throughput of 40 gigaFLOPs, the model is computationally efficient and therefore suitable for real-time application scenarios.

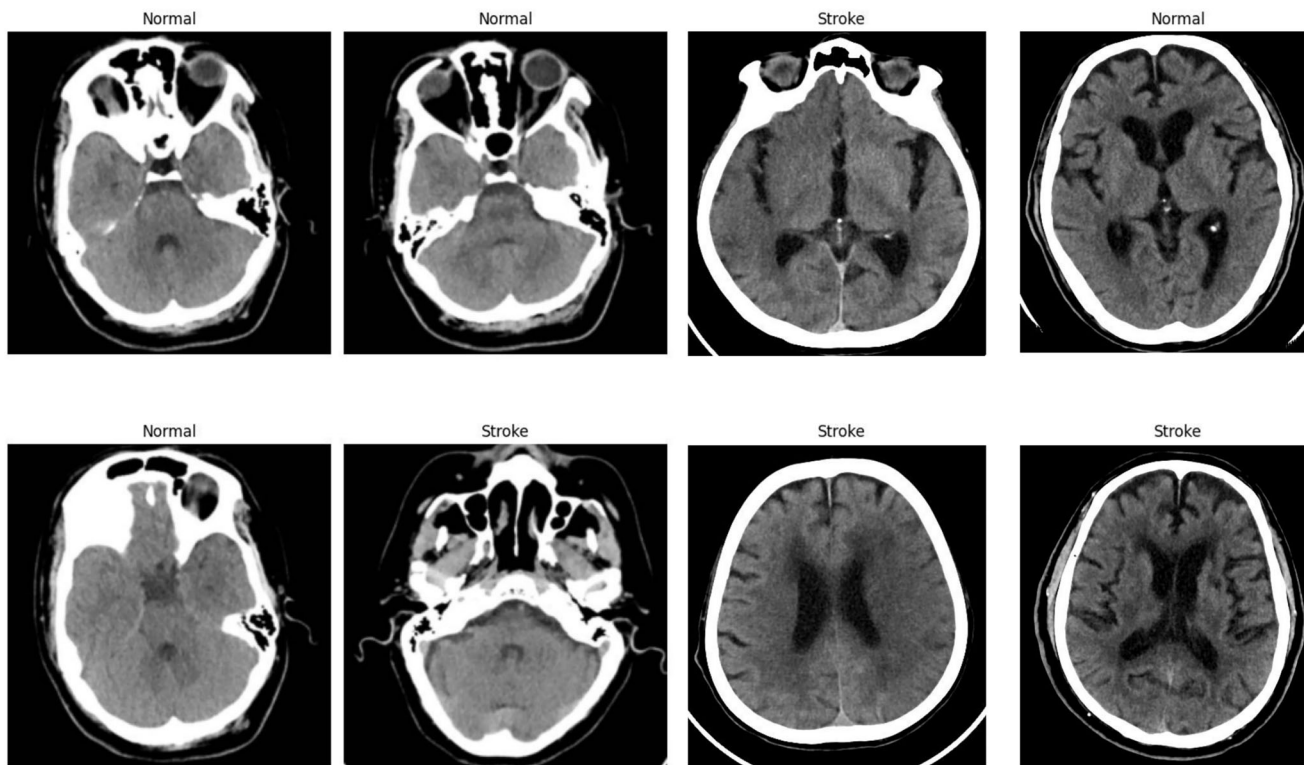


Fig. 1 Few samples CT images from public and private datasets

The proposed study aims to develop advanced artificial intelligence stroke detection techniques and practical solutions to assist clinicians in low-resource settings. Deep learning algorithms are commonly used in medical imaging to support early detection and classification of strokes when applied to computed tomography (CT) images.

The organization of the study is as follows. The literature on stroke classification and the application of deep learning in medical imaging is reviewed in “[Literature review](#)”. The materials and methods, preprocessing steps and data augmentation are described in “[Materials and methods](#)”. “[Proposed framework](#)” describes the architecture and optimization procedure of the proposed StrokeFuse-AttnNet model. The design and optimization of the model are detailed in “[Training and optimization](#)”; experimental results are discussed in “[Experimental results](#)” and the results are presented and discussed in “[Discussion](#)”. “[Conclusion](#)” concludes the investigation and suggests directions for future studies.

Literature review

Recent studies aim to evaluate different architectures, feature extraction methods and image pre-processing techniques to further increase diagnostic accuracy. In general, a deep learning model using ResNet [20] is used to detect early ischaemic

stroke in non-contrast CT images (NCCT). Research shows that ResNet can identify early signs of anaerobiosis due to its ability to process complex, high-dimensional data. Classification of CT images into stroke subtypes is performed using the ResNet-50 architecture, achieving 98.72% accuracy. Effective hyperparameter tuning enables this performance, demonstrating the value of the architecture in stroke detection [21]. High stroke detection performance is achieved by the scalable architecture and processing capacity of the industry-leading EfficientNet-B0 solution. Patel et al. (2023) achieve a stroke classification accuracy of 97% with EfficientNet-B0, while previous approaches achieve lower performance [22]. A benchmark test was conducted with EfficientNet-B0 and other architectures ResNet101, VGG19 and GoogleNet—where the highest accuracy of 97.93% was achieved on the CT dataset, demonstrating the optimal performance of the model [23].

A computer-aided diagnosis system employs convolutional neural networks to classify CT scans and distinguish between normal cases, ischaemic strokes and haemorrhages. Data augmentation includes horizontal flipping with real-time transformations and early stopping techniques are used to prevent overfitting. The model achieved an accuracy of 79%, demonstrating that augmentation strategies enhance generalization abilities [13]. The early stroke detection system applies a genetic algorithm for feature selection and

then uses bidirectional long short-term memory (BiLSTM) to classify early strokes. Integrating these two methods resulted in an accuracy of 96.5%, demonstrating their effectiveness in stroke diagnosis [14].

Hybrid CNN and attention-based mechanisms are increasingly used to diagnose strokes. Abumihsan et al. (2025) proposed a CBAM-enhanced hybrid CNN model that demonstrated improved localisation of ischaemic strokes on CT scans [24]. Similarly, Zhou et al. (2024) used lightweight attention-based CNNs in mobile gait analysis systems for post-stroke rehabilitation, emphasizing the role of attention in real-time applications [25]. Aksoy et al. (2024) optimized stroke classification by combining pre-trained CNN-LSTM models with attention modules to process multi-modal clinical inputs [26]. Meanwhile, Vindas et al. (2022) proposed a CNN transformer for brain embolism detection that, while not stroke-specific, is an effective multi-feature attention strategy for neuroimaging applications [27].

A hybrid CNN-RNN system was developed, along with pre-processing techniques for data segmentation and image concatenation, to detect intracranial haemorrhages in NCCT scans. This approach achieved significant performance improvements, with the AUC-ROC increasing from 0.854 to 0.966 with pre-processing and the model [12]. These results are consistent with the proposed approach, which incorporates advanced pre-processing and normalisation techniques. The results of the SqueezeNet and MobileNet models are combined by feature fusion and then processed with CatBoost for stroke classification. Using this approach, a significant accuracy of 99.1% was achieved, demonstrating the effectiveness of combining deep learning architectures with traditional machine learning techniques [28]. While machine learning (ML) and deep learning (DL) systems have been used for stroke diagnosis, their methods, data requirements and performance differ. SVMs and decision tree machine learning models are generally effective for structured tabular clinical data, especially when datasets are small and features are expert-engineered. These models are easier to interpret than others but fail to capture the intricate spatial patterns of neuroimages and rely heavily on feature quality. In contrast, deep learning models can easily diagnose medical images. Convolutional Neural Networks (CNNs) are well suited to this task. These models perform well in stroke classification, haemorrhage detection and infarct localisation and automatically extract deep, hierarchical feature representations from CT or MRI images. When manual feature organization is infeasible, such as with large imaging datasets, deep learning models are valuable. They typically require more data, greater processing power and careful regularization to prevent overfitting.

In addition to deep learning architectures, many machine learning (ML) paradigms have proven useful for modelling complex and nonlinear patterns in various scientific and

engineering applications. Classical and modern neural network-based ML models have been successfully applied to nonlinear system modelling and prediction tasks, demonstrating high representational power and adaptability across domains such as engineering, agriculture and complex system analysis [29–31]. In addition to deep learning for images, these studies demonstrate the flexibility of neural architectures in other problem domains and encourage hybrid designs that exploit complementary feature representations.

Gaussian process regression has received considerable attention because its probabilistic nature allows quantification of predictive uncertainty. This makes it a valuable approach in safety-critical and data-limited applications [32–34]. Although GPR models are often computationally constrained for high-dimensional image data, their focus on uncertainty modelling provides important conceptual insights that complement attention mechanisms used in deep learning, especially for improving model reliability and interpretability in medical diagnosis.

Graphical and probabilistic models provide a powerful framework for learning structured dependencies and causal relationships among variables [35–37]. These models emphasize explainability and structured reasoning, which are increasingly important in clinical decision-support systems. Although graphical techniques are less suited to direct raw image processing, their principles of dependency modelling align with the objective of attention mechanisms, which selectively emphasise informative regions and feature interactions.

Ensemble and composite learning methods further enhance robustness and generalization by integrating multiple learners, thereby reducing variance and improving stability across heterogeneous datasets [38–40]. These strategies are conceptually related to the proposed hybrid feature fusion framework, in which complementary representations extracted from multiple backbone networks are combined to improve discrimination performance and resilience to dataset variability.

These broader machine learning paradigms strengthen the rationale for creating StrokeFuse-AttnNet. The feature fusion strategy incorporates the ensemble learning principle in its design. The interpretation of the self-attention mechanism can be readily aligned with structured dependency modelling in graphical approaches. Additionally, the refinement of salient features corresponds to uncertainty-aware perspectives in Gaussian process modeling. This method is presented as part of a unified framework for achieving robustness, interpretability, and high performance across a broad range of medical imaging tasks, which is an important aspect of the wider machine learning ecosystem.

Recent advances in medical image analysis have explored various hybrid and attention-based deep learning architectures, including dual-path networks, attention-gated CNNs

and feature fusion frameworks, to enhance representation learning and localization performance. Consequently, such methods provide improved accuracy for stroke classification and other neuroimaging-related tasks. However, most existing models require evaluation on a single dataset, use complex multi-head attention mechanisms with high computational cost and have limited interpretability for clinical deployment.

In addition, many previous studies have focused separately on global semantic feature extraction or local texture modelling, which may affect the generalisability of heterogeneous clinical datasets. Interpretable models are typically modified after the architecture, although this involves the topical relevance of the feature. These limitations highlight the need for a lightweight, interpretable and generalizable framework that integrates complementary feature representations while remaining suitable for real-world clinical environments.

The StrokeFuse-AttnNet is a new framework featuring a hierarchical approach to feature fusion, unlike existing hybrid and attention-based designs. It combines global features from ResNet50 with fine-grained, reusable features from DenseNet121, then refines the fused features using a lightweight self-attention mechanism. This design enables explicit modelling of complementary feature dependencies with high efficiency. Furthermore, the proposed framework adopts a targeted fine-tuning strategy and is systematically evaluated across both public and private datasets, demonstrating improved generalization and interpretability. This study shows that the introduction of attention alone is not novel, but that elaborate fusion and attention integration can effectively circumvent important robustness and clinical trust issues.

Materials and methods

Pre-trained CNN models

Convolutional Neural Networks are deep learning models widely used for classification and feature extraction tasks [41]. By using pre-trained models, it is possible to leverage features learned from large datasets such as ImageNet to improve performance and efficiency. In this study, feature extraction was performed using ResNet50 and DenseNet121. The architectures of these models contributed to their effective performance on the dataset.

ResNet50

ResNet50 uses a 50-layer residual connection to address the problem of vanishing gradients. Skip connections enable some layers to bypass certain operations, making it easier

to learn deeper patterns without disrupting the network's operations. ResNet50 utilises a combination of 3×3 and 7×7 convolutional kernels, with an input image size of $224 \times 224 \times 3$. To enable ResNet50 to accept single-channel grayscale CT images, the first convolutional layer was modified. The final layers were also adjusted to provide binary stroke classification outputs for this model. In this fine-tuning process, although the first layers were adapted to this dataset, the features learned in those layers are retained.

DenseNet121

DenseNet121 is a deep network with 121 layers that employs complex interconnections. DenseNet differs from traditional CNNs as it enables the transfer of feature maps to higher layers [42]. The architecture promotes feature reuse and reduces redundancy. DenseNet121 uses 7×7 kernels in its first convolutional layer and processes input images of size $224 \times 224 \times 3$. The first convolutional layer receives grayscale CT images as input and the final classification layers are modified to detect strokes. The model accurately identifies complex spatial patterns in images through dense connections and feature reuse. The proposed model is paired with the pre-trained models ResNet50 and DenseNet121. ResNet50 extracts hierarchical features, while DenseNet121 reuses features. To enhance spatial perception and classification ability, a self-attention mechanism is incorporated into the architectures.

Datasets

The proposed model was trained and tested using two distinct datasets: a publicly available dataset and a private dataset. The proposed model is robust as it includes both normal and stroke case CT brain images.

Public Dataset: The public dataset used in this study was obtained from an openly available brain CT stroke repository [43]. The dataset contains non-contrast CT brain images acquired from various sources and scanners. The images were obtained according to real-life clinical protocols. There are natural variations in imaging resolution and scanner manufacturers. Expert radiologists have labelled the images as normal or stroke. All images are in greyscale. The scanners and acquisition parameters are not always available in detail, so this inherent heterogeneity makes it suitable for assessing the generalisability of deep learning methods.

Private Dataset: A private dataset comprising 6,228 non-contrast CT brain images was collected from the Chulalongkorn Stroke Centre. Before analysis, all CT scans were anonymised and the data were used solely for research in accordance with the institution's policies and guidelines. The study protocol was approved for retrospective analysis

Table 1 Distribution of classes in public and private datasets

Dataset	Stroke cases	Normal cases	Total
Public dataset	950 (37.98%)	1551 (62.02%)	2501
Private dataset	1868 (29.99%)	4360 (70.01%)	6228
Total	2818	5911	8729

and no identifiable patient information was accessible to the researchers.

Board-certified neuroradiologists examined each CT scan and labelled it as normal or stroke. Information on patient demographics and clinical metadata, including age distribution, stroke subtype and time since onset, was not consistently available for all cases and was therefore excluded from the analysis. This limitation is acknowledged and may affect subgroup-level interpretability. However, the dataset reflects real-world clinical variability and supports evaluation of the model's robustness in routine diagnostic settings.

As shown in Table 1, this dataset comprises 8729 CT images, with 4360 stroke-positive and 4369 normal cases. Overall, the dataset is fairly balanced. However, there are different distributions for public and private images in the dataset. This means the training and testing splits are somewhat imbalanced. When this bias exists, the learning may also become biased and be less sensitive to minority cases. To address class imbalance, the training data were balanced using the Synthetic Minority Over-sampling Technique (SMOTE). In this study, SMOTE was applied in the feature space rather than directly in the image space. Specifically, synthetic samples were generated from deep feature representations extracted from the training data by the backbone networks. Because applying SMOTE directly to pixel values creates unrealistic artefacts in medical images, our implementation in feature space prevents this while still improving class balance during training. Although SMOTE enhances class distribution, it can generate synthetic samples that do not accurately reflect variations in pathology. Furthermore, it may cause decision boundaries to become over-smoothed. To reduce this risk, SMOTE was applied conservatively and combined with extensive on-the-fly data augmentation to maintain feature diversity. Alternative imbalance-handling techniques using a class-weighted loss or focal loss were also considered. However, we chose SMOTE for its simplicity, effectiveness on moderate imbalance and because it does not introduce additional hyperparameters or training instability.

Data augmentation

An effective data augmentation strategy was used on the fly during training to reduce overfitting and improve general-

ization. During each epoch, image inputs were randomly transformed using a fixed set of augmentations with explicitly defined parameters. Specifically, images were randomly rotated within a range of $\pm 15^\circ$, horizontally flipped with a probability of 0.5 and randomly cropped and resized with a scale factor from 0.8 to 1.0. Brightness and contrast were adjusted within a range of $\pm 20\%$ to simulate illumination variability and random erasing was applied with a probability of 0.2 to mimic partial occlusions commonly observed in clinical CT scans. The transformations mentioned above were applied dynamically during training, rather than creating a static augmented dataset. This enabled the model to encounter different versions of each class 100 times per training epoch without increasing the memory footprint. All CT images were resized to 224×224 pixels to ensure compatibility with the input dimensions of the pre-trained ResNet50 and DenseNet121 architectures. The following normalization equation was then applied to standardize the input images.

$$x_{\text{normalized}} = \frac{x - \mu}{\sigma}, \quad (1)$$

where $\mu = [0.5]$ and $\sigma = [0.5]$, reflecting the grayscale image of the input data. These augmentations improved performance on the testing dataset with robust training.

Proposed framework

The proposed model begins with feature extraction and concludes with final classification. We propose an integrated multi-stage process for classifying strokes from CT images. This framework identifies useful features and redundancies in the dataset.

The model uses pre-trained, optimized CNNs ResNet50 and DenseNet121 for CT images. ResNet50 is effective for feature extraction due to its residual connections, while DenseNet121 performs well because of its dense connectivity. The outputs of these models are concatenated to form a unified feature representation. A self-attention mechanism is then applied to the concatenated features to highlight the most important spatial and channel information, thereby improving classification accuracy. The flow diagram of the proposed system is shown in the Fig. 2.

Feature extraction

The pre-trained models ResNet50 and DenseNet121 were modified to process single-channel grey images. The weights of both models were initialized with ImageNet. The first layers of all models were retained, as they have the potential

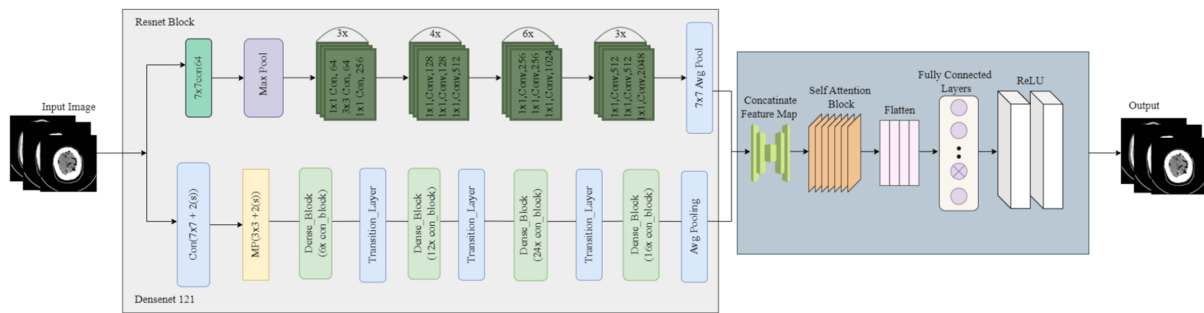


Fig. 2 Framework of the proposed model

to extract useful features. We only modified the last layers and fine-tuned them. ResNet50 extracts feature maps of size $\mathbb{R}^{N \times 2048}$ and DenseNet121 generates feature maps of size $\mathbb{R}^{N \times 1024}$, where N denotes the batch size. To evaluate the quality of the extracted features, t-SNE (t-distributed Stochastic Neighbour Embedding) was applied to the features generated by the StrokeFuse-AttnNet model for both the public and private datasets. The t-SNE visualizations shown in Figs. 3 and 4 illustrate the separability of the Normal and Stroke classes in feature space. In both ResNet50 and DenseNet121, feature extraction is performed from the final convolutional block immediately before the global average pooling layer. The feature maps produced by the final convolutional stage are retained and organized to form high-dimensional feature vectors. This approach preserves spatially rich representations while capturing high-level semantics. ResNet50 has a deeper architecture with residual connections, allowing it to extract features with global context. In contrast, DenseNet121 has dense feature reuse in the intermediate layers, enabling it to extract features with more localised and fine-grained information. The dual use of these features as both global and local feature extractors in the proposed hybrid framework.

For the public dataset in Fig. 3, the t-SNE diagram shows clear clustering of the Normal and Stroke classes, with some overlap. The features from the data demonstrate moderate discriminative power, which corresponds with the high score of 98.54%. In the private dataset in Fig. 4, there is clear overlap between the Normal and Stroke classes, illustrating the difficulty in distinguishing these classes. This overlap corresponds to the lower accuracy of 91% for the Stroke class, as shown in Table 4. The results show that using self-attention refinement methods separates learned features more effectively, leading to higher accuracy.

t-SNE preserves the local neighbourhood structure in a low-dimensional space for visualisation, without affecting class separation in the original high-dimensional feature space. Therefore, the measured overlap in two-dimensional embeddings does not indicate that the learned features are less discriminative. The high classification performance on

the private dataset suggests that the model captures complex, nonlinear feature interactions that are not fully preserved under dimensionality reduction. We also recognize the possibility that dataset-specific acquisition characteristics or scanner-related variations may contribute to model predictions. Future work will involve institutions conducting artefact-controlled experiments and validating explainability against expert annotations to increase the robustness of measurements.

Feature fusion strategy

ResNet50 and DenseNet121 were used for feature extraction and feature concatenation. The concatenated feature vector, $\mathbf{F}_{fusion} \in \mathbb{R}^{N \times (2048+1024)}$, combines global and local features from the input images. This hybrid strategy preserves the complementary advantages of both models. Although concatenation introduces some unnecessary and irrelevant data, this is mitigated by self-attention. Before applying the self-attention mechanism, the concatenated feature vector is normalized and scaled to ensure numerical stability and balanced contributions from both backbone networks. Specifically, batch normalization is applied to the fused feature representation to standardize feature distributions and reduce internal covariate shift. Additionally, a learnable scaling parameter adaptively adjusts the magnitude of the fused features before attention, preventing high-variance feature channels from deeper layers of ResNet50 from dominating. This normalization and scaling step improves convergence stability and enables the attention mechanism to focus on semantically meaningful feature interactions.

Self-attention mechanism

The self-attention mechanism is used to refine the fused feature representations by selectively emphasizing the most informative spatial regions for stroke classification [44]. In the proposed framework, self-attention is applied to spatially tokenized feature maps extracted from the final convolutional blocks of ResNet50 and DenseNet121 before global aver-

Fig. 3 t-SNE visualization of the feature space for the public dataset. Blue and orange represent the classes Normal and Stroke respectively

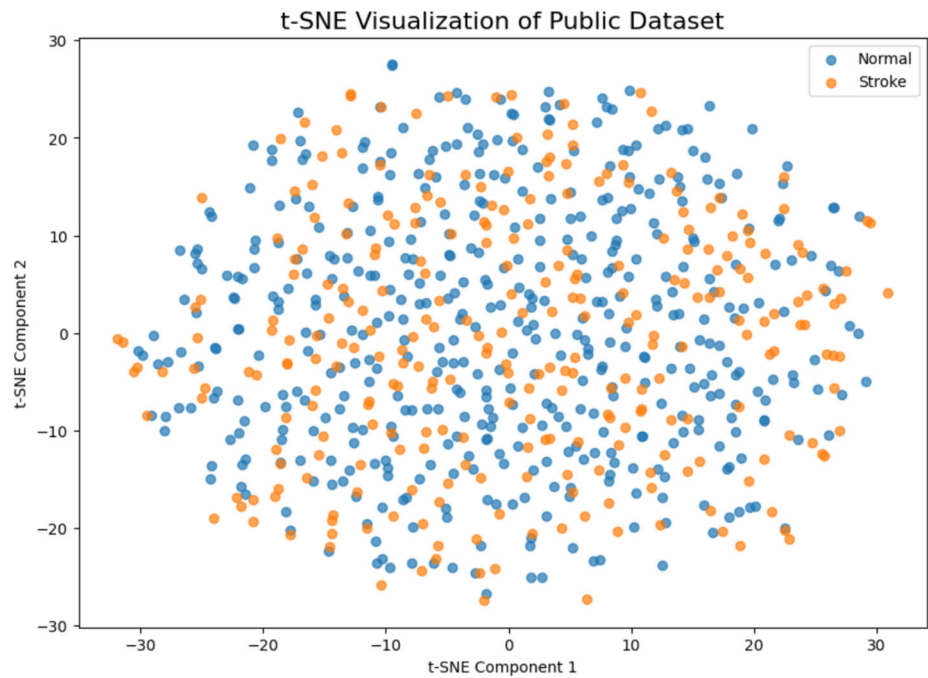
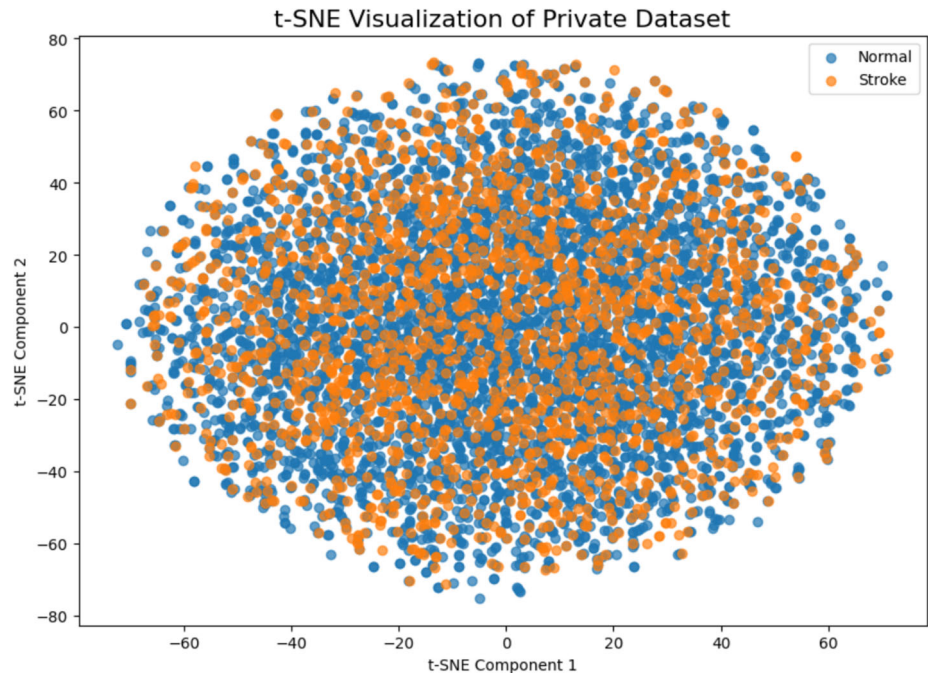


Fig. 4 t-SNE visualization of the feature space for the Private data set



age pooling. Specifically, the spatial feature maps from both backbones are reshaped into a sequence of tokens and concatenated along the channel dimension to form a unified representation.

Scaled dot-product self-attention is then used to model long-range dependencies among these spatial tokens, enabling the network to capture contextual relationships across different anatomical regions of the brain. The resulting attention weights highlight discriminative stroke-related regions while suppressing irrelevant or noisy activations. The attention-

refined tokens are subsequently aggregated into a compact feature vector for classification. This spatially aware attention design improves both classification accuracy and interpretability and is consistent with the spatial feature visualizations shown in Figs. 5 and 6.

Training and classification

Fully connected layers are then used to classify the refined features into two classes. Dropout layers are included to pre-

vent overfitting and final classification is performed using a sigmoid activation function. The model is trained with binary cross-entropy loss and the Adam optimizer. To ensure stable convergence, a stepwise learning rate scheduler is applied. Training is conducted with automatic mixed precision (AMP) on graphics processing units (GPUs). The proposed model incorporates ResNet50, DenseNet121 and a self-attention mechanism for effective visualization and interpretation of stroke classification using CT.

Model architecture

The proposed model utilises two CNNs pre-trained on ImageNet: ResNet50 and DenseNet121. These models are modified for single-channel grayscale images and adapted for binary stroke classification. ResNet50 facilitates efficient hierarchical feature learning through its residual connections, while DenseNet121 employs densely connected layers to enhance feature reuse and efficient gradient flow. Figure 5 shows a detailed flowchart illustrating the process from data preprocessing to classification. The feature maps extracted by both models are concatenated into a single feature representation.

$$F_{\text{fusion}} = F_{\text{ResNet50}} \oplus F_{\text{DenseNet121}}, \quad (2)$$

where $F_{\text{ResNet50}} \in \mathbb{R}^{N \times 2048}$ and $F_{\text{DenseNet121}} \in \mathbb{R}^{N \times 1024}$ represent the feature maps, N is the batch size and \oplus denotes concatenation. The resulting feature vector $F_{\text{fusion}} \in \mathbb{R}^{N \times 3072}$ captures complementary global and local features.

A self-attention mechanism is then used to refine the fused feature representation. It estimates the attention weights based on learning dependencies between the feature elements, making the important features more important for denoising. Finally, the self-attention mechanism computes the query, key and value matrices as follows:

$$Q = W_q \cdot F_{\text{fusion}}, \quad K = W_k \cdot F_{\text{fusion}}, \quad V = W_v \cdot F_{\text{fusion}}, \quad (3)$$

where W_q , W_k and W_v are learnable weight matrices for the query, key and value transformations. The attention weights are calculated using the scaled dot product mechanism:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V, \quad (4)$$

The scaled dot-product self-attention formulation was chosen for its computational efficiency and numerical stability when applied to high-dimensional fused feature vectors. The method facilitates global dependency modelling with less overhead compared to multi-head attention, which increases parameters and computational cost. Channel-based attention

mechanisms do not explicitly capture long-range dependencies from the fused feature space but mainly recalibrate feature importance. Because it is an attention mechanism that seamlessly achieves expressiveness, efficiency, and real-world applicability.

where d_k is the dimensionality of the key vectors. The refined feature representation is then calculated as:

$$F_{\text{refined}} = \gamma \cdot \text{Attention}(F_{\text{fusion}}) + F_{\text{fusion}}, \quad (5)$$

where γ is a learnable scalar parameter that regulates the contribution of the attention-refined features relative to the original fused feature representation.

The refined feature vector F_{refined} is then passed through fully connected layers for binary classification. During training, to reduce overfitting, a dropout layer with a rate of 0.5 is applied. Finally, the classification probability is generated using a sigmoid activation function.

$$y_{\text{output}} = \sigma(W \cdot F_{\text{refined}} + b), \quad (6)$$

where W and b are the weights and biases of the output layer and σ is the sigmoid activation function.

Table 2 provides a detailed breakdown of the trainable parameters used in the proposed StrokeFuse-AttnNet architecture.

The Fig. 6 shows the workflow of the proposed model, including the extracted features, attention and predictions. This visualization presents examples of feature extraction by ResNet50 and DenseNet121, shown in Fig. 6b, c, respectively. The hybrid fusion of the extracted features is depicted in Fig. 6d, while the attention-based refinement process is highlighted in Fig. 6e. Figure 6f–h show Grad-CAM overlays and saliency maps, which highlight the discriminative regions contributing to stroke detection. These illustrations provide insight into the model's operation.

Proposed algorithm

Algorithm 1 describes a systematic approach to stroke classification that incorporates self-attention mechanisms for feature extraction and classification. The model achieves high accuracy and effective generalization in stroke detection through the inclusion of pre-processing, feature fusion and feature refinement. The proposed model operates in several stages. First, the provided CT images are preprocessed and standardized using a mean and standard deviation of 0.5. During this phase, an advanced data augmentation protocol is applied. this protocol includes random rotation, flipping, cropping, brightness and contrast adjustment and random deletion.

Table 2 Detail of trainable parameters in StrokeFuse-AttnNet

Component	Trainable parameters (millions)
ResNet50 (fine-tuned layers only)	~ 18.1
DenseNet121 (fine-tuned layers only)	~ 11.2
Self-Attention Module	~ 1.1
Fully connected classifier	~ 1.6
Total trainable parameters	~ 32.0

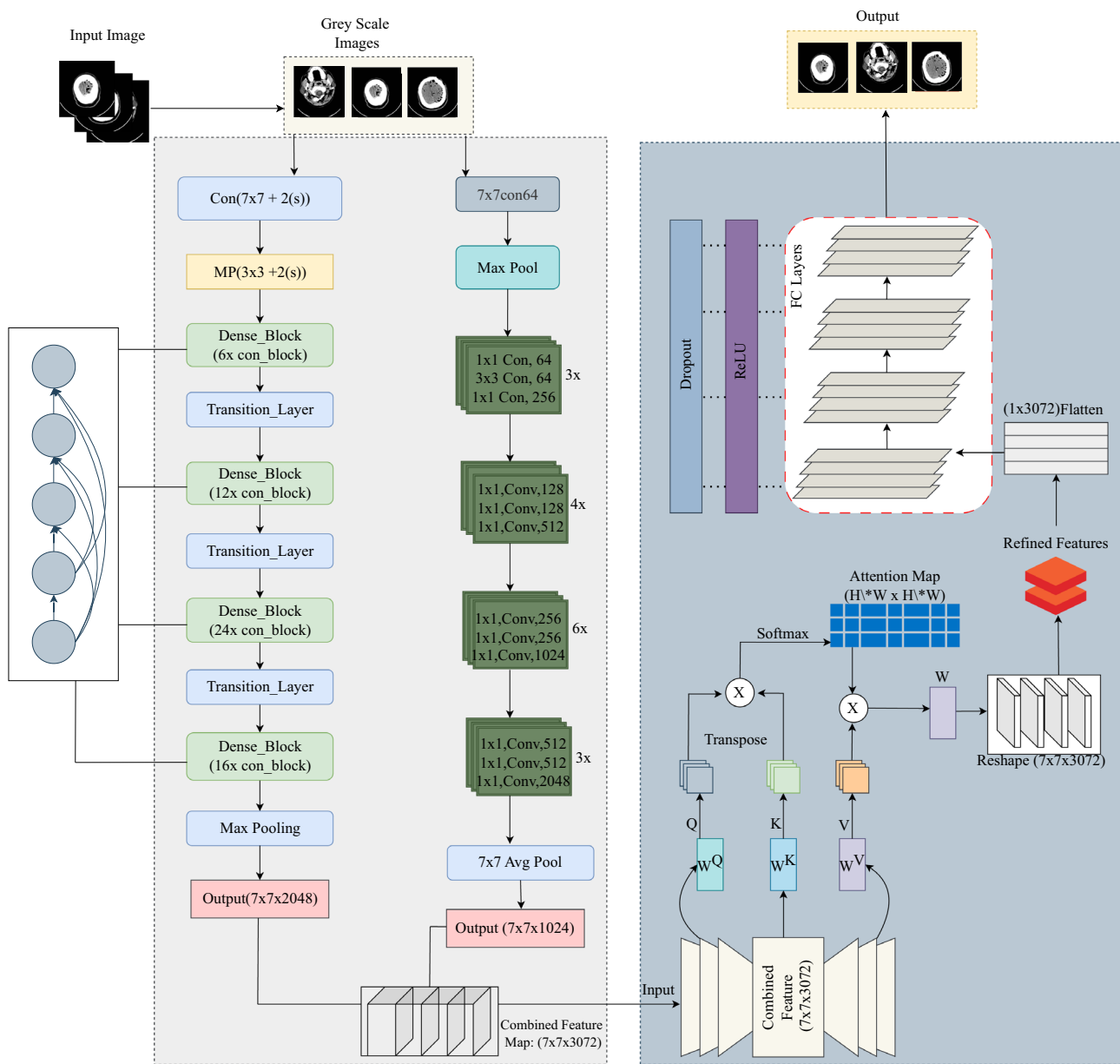


Fig. 5 Architecture of the proposed StrokeFuse-AttnNet model. Grayscale CT images are processed in parallel using DenseNet121 and ResNet50 to extract complementary local and global features. The

resulting feature maps are concatenated and refined through a self-attention mechanism before being passed to fully connected layers for final stroke classification

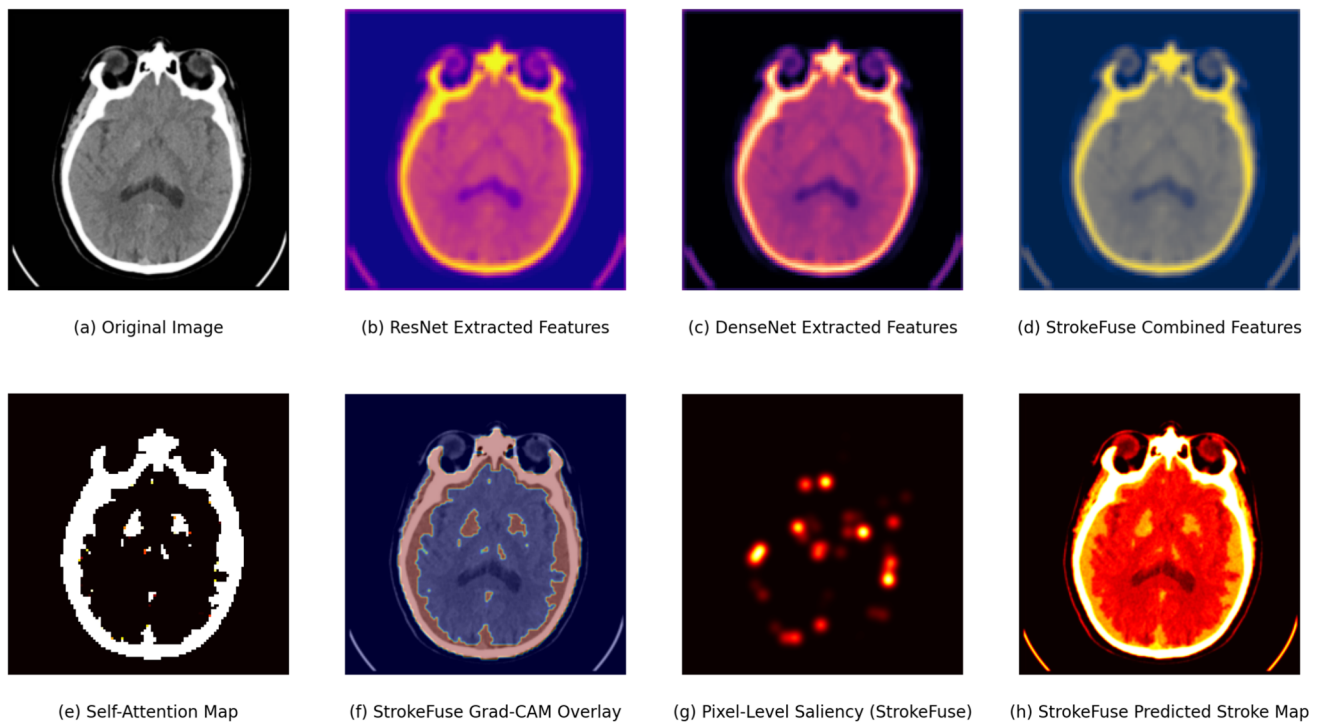


Fig. 6 Visualization of StrokeFuse-AttnNet operation: **a** original CT image; **b** features extracted by ResNet; **c** features extracted by DenseNet; **d** combined features using hybrid feature fusion; **e** self-

attention; **f** Grad-CAM overlay highlighting important regions; **g** pixel-level saliency visualization; **h** predicted stroke map showing the final model output

The next phase, feature extraction, employs two deep learning architectures: ResNet50 and DenseNet121. The feature maps from the backbone output are fused to obtain a fine-grained common feature representation. After feature extraction, the concatenated features are processed by self-attention. This mechanism calculates the correlation of the concatenated features and assigns a weight to each feature, focusing on important regions and discarding less relevant ones. The features extracted from the output are then passed to fully connected layers to classify the stroke.

Training and optimization

For stable convergence, robustness, and efficient deployment of StrokeFuse-AttnNet, specify the training strategy and optimization settings in this section. The network was trained end-to-end using a supervised learning framework, with all experiments conducted under identical conditions to ensure fair evaluation.

The model was optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} , providing a suitable balance between convergence speed and training stability. The model uses the StepLR learning rate scheduler, which reduces the learning rate by a factor of 0.5 every five epochs

Algorithm 1 Workflow of StrokeFuse-AttnNet

- 1: **Input:** Dataset \mathbf{D} , batch size $N = 16$, learning rate $\eta = 1 \times 10^{-4}$, epochs $E = 100$
- 2: **Output:** Binary classification (Stroke / Normal)
- 3: **Step 1: Preprocessing and Augmentation**
- 4: Resize images to 224×224 and normalize: $x_{\text{norm}} = \frac{x-0.5}{0.5}$
- 5: Apply random rotations, horizontal flipping, cropping, brightness and contrast adjustments and random erasing.
- 6: **Step 2: Feature Extraction and Fusion**
- 7: Fine-tune the last 10 layers of ResNet50 and DenseNet121
- 8: Extract features: $\mathbf{F}_{\text{ResNet50}} \in \mathbb{R}^{N \times 2048}$, $\mathbf{F}_{\text{DenseNet121}} \in \mathbb{R}^{N \times 1024}$
- 9: Fuse features: $\mathbf{F}_{\text{fusion}} = \mathbf{F}_{\text{ResNet50}} \oplus \mathbf{F}_{\text{DenseNet121}}$
- 10: **Step 3: Attention-based Refinement**
- 11: Compute: $Q = W_q \mathbf{F}_{\text{fusion}}$, $K = W_k \mathbf{F}_{\text{fusion}}$, $V = W_v \mathbf{F}_{\text{fusion}}$
- 12: Apply scaled dot-product attention: $\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
- 13: Refine features: $\mathbf{F}_{\text{refined}} = \gamma \cdot \text{Attn}(\mathbf{F}_{\text{fusion}}) + \mathbf{F}_{\text{fusion}}$
- 14: **Step 4: Classification**
- 15: Apply dropout: $\mathbf{F}_{\text{dropout}} = \mathbf{F}_{\text{refined}} \cdot \text{Bernoulli}(1 - p)$, $p = 0.5$
- 16: Predict output: $y_{\text{out}} = \sigma(W\mathbf{F}_{\text{dropout}} + b)$
- 17: **Step 5: Training**
- 18: Compute loss: $\mathcal{L}_{\text{BCE}} = -\frac{1}{M} \sum_{i=1}^M [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$
- 19: **for** $e = 1$ to E **do**
- 20: Update parameters using Adam optimizer
- 21: **if** $e \bmod 5 = 0$ **then**
- 22: Adjust learning rate: $\eta \leftarrow 0.5\eta$
- 23: **Step 6: Return** y_{out}

to avoid local minima. For the binary classification task, the objective function is the binary cross-entropy loss.

Training was conducted for 100 epochs with a batch size of 32, chosen to balance GPU memory usage and gradient stability. Dropout at a rate of 0.5 was applied to the fully connected layers to reduce overfitting. To retain the generic visual representations learned from ImageNet, the earlier layers of the pre-trained ResNet50 and DenseNet121 backbones were frozen, while the final convolutional blocks, self-attention module and classifier layers were fine-tuned on the stroke datasets.

To enhance robustness, extensive on-the-fly data augmentation was applied during training and class imbalance was addressed using feature-space SMOTE, as described earlier. All experiments were run on an NVIDIA RTX 4070 GPU and mixed-precision training was used to improve computational efficiency. The proposed model is a near real-time clinical model due to its favorable balance between accuracy, generalization and computational cost.

Training protocol

The Binary Cross-Entropy (BCE) loss function trains the model for binary classification.

$$L_{\text{BCE}} = -\frac{1}{M} \sum_{i=1}^M [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)], \quad (7)$$

where M is the batch size, y_i is the ground truth label and p_i is the predicted probability.

An Adam optimizer updates the model parameters according to a specified rule, which is used for optimization.

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t + \epsilon}}, \quad (8)$$

where η represents the learning rate, m_t and v_t denote the first and second moment estimates respectively and ϵ is a small constant that ensures numerical stability.

For stable convergence, we use a StepLR scheduler that reduces the learning rate by a factor of 0.1 every 5 epochs.

$$\eta_{\text{new}} = \eta_{\text{current}} \cdot \gamma, \quad (9)$$

where $\gamma = 0.5$ is the decay factor.

Regularization is applied through dropout during training, with a rate of 0.5:

$$F_{\text{dropout}} = F_{\text{refined}} \cdot \text{Bernoulli}(1 - p), \quad (10)$$

where $p = 0.5$ represents the dropout probability.

Experimental setup

An environment equipped with an NVIDIA RTX 4070 graphics processor and an Intel Core i9 processor provides sufficient computational power to train deep learning models. This environment was used for experimentation. The implementation was in Python using the PyTorch framework and was trained and run using CUDA. The dataset for the study consisted of enhanced CT scans organized in folders. A hold-out evaluation protocol was employed, with 70% of the data used for training and 30% reserved for independent testing. Model training and evaluation were strictly separated to prevent data leakage. This split was applied to the combined datasets for an overall robustness assessment on heterogeneous data sources. However, it is not intended as a strict cross-institutional generalization experiment, which will be assessed in future work.

The model was trained for 100 epochs, during which stable convergence was observed. A batch size of 32 was chosen to balance memory efficiency and gradient stability during backpropagation. The use of an Adam optimizer with a learning rate scheduler enhanced training robustness. These aspects ensured stable optimization and consistent performance during training and testing of both the Grover model and the Naïve model. The StrokeFuse-AttnNet model was enhanced using a range of high-performance devices and software platforms in a dedicated experimental setup. This facility ensured consistent experimental results. Consequently, the training and evaluation workflows were integrated.

Hyperparameters

The StrokeFuse-AttnNet model's hyperparameters, learning rate, batch size, number of epochs, optimisers and data augmentation methods determine the efficiency and performance of stroke classification training. Hyperparameter selection was conducted using a structured sensitivity analysis rather than an exhaustive grid search or Bayesian optimization strategy. Key hyper-parameters, including learning rate, batch size and dropout rate, were varied individually while keeping other parameters fixed and their impact on validation accuracy and AUC was monitored. This approach was adopted to ensure stable convergence and generalization while maintaining computational efficiency, given the scale of the datasets and the depth of the proposed architecture.

The learning rate determines the path of gradient descent. The batch size affects training stability and resource utilization and the Adam optimizer was used to adjust the learning rate for the model trained over 100 epochs. A StepLR scheduler reduced the learning rate by 50% every 5 epochs to avoid overfitting. The dropout rate used was 0.5. The generalization of the model was improved by data

augmentation methods such as rotation, flipping, cropping, changes in brightness/contrast and random dropouts. Table 3 summarizes the hyperparameter used to train the StrokeFuse-AttnNet model.

Evaluation metrics

In this study, the model's performance is measured using evaluation metrics including accuracy, precision, recall, F1 score and AUC-ROC. These metrics enable assessment of the model's stroke classification capabilities. The definitions are as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad (11)$$

where TP and false counts by FP and FN . The F1 score, defined as the harmonic mean of precision and recall, is calculated as follows:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (12)$$

The AUC-ROC curve is used to measure the ability of the model to distinguish strokes from normal cases. We calculate the AUC by calling `roc_curve` and `auc` from scikit-learn, which calculate the area under the ROC curve.

$$\text{AUC} = \int_0^1 TPR(x) dx, \quad (13)$$

where $TPR(x)$ is the true positive rate as a function of the false positive rate (x). Beyond conventional classification metrics, recent studies have highlighted the importance of complementary performance measures that capture normalized error magnitude, information consistency and predictive reliability, particularly for nonlinear modelling and clinical decision support systems. Relative Root Mean Square Error (RRMSE) has been widely used as a normalized error metric to enable fair performance comparison across models and datasets with different scales, especially in complex nonlinear prediction tasks [33, 45]. Such normalized error-based measures provide additional insight into model stability beyond accuracy-based evaluation.

Metrics based on information theory, such as Normalised Mutual Information (NMI), can quantify the agreement between predicted labels and ground truth. They also consider the interdependence of distributions [46]. Alternative methods such as the Brier score and Expected Calibration Error (ECE) are gaining traction for evaluating the quality of probabilistic predictions, which is relevant for clinically risk-aware deployments. Although the study primarily focuses

on classification performance, these metrics provide valuable perspectives for future uncertainty-aware extensions of the proposed framework. Performance metrics with confidence intervals to support hypothesis testing and account for statistical variability. This study has discussed that bootstrap confidence intervals are suitable for estimating metric uncertainty. Furthermore, it recommends using statistical tests such as DeLong's test for AUC and McNemar's test for classification accuracy for rigorous model comparison. The importance of these analyses increases when performance differences are small, while future versions will use formalised statistical testing over many runs.

Experimental results

Results on public dataset

The Stroke-FuseAttnNet model achieves high accuracy in classifying stroke and non-stroke CT images from the dataset. The ROC curve, confusion matrix, and classification report are used to evaluate the effectiveness of the architecture.

Figure 7 shows the model's ability to discriminate between stroke and normal cases, with an area under the curve of 0.98 and high sensitivity and specificity. An AUC value indicates that the model is valid for detecting pathological changes associated with stroke. Figure 8 shows the results of the model's predictions. Among the normal cases, 460 were correctly identified, while 8 were misclassified as stroke. Of the stroke cases, 277 were accurately categorized and 5 were incorrectly predicted as normal. The StrokeFuse-AttnNet model attains an accuracy of 98.27%. The model also demonstrates strong class-specific and overall performance metrics. For normal cases, the precision is 97.19%, while for stroke cases it is 98.23%. The F1 score is 97.70% and the specificity is 98.29%, demonstrating the effectiveness and generalization of the model in the presence of class imbalance. Table 5 summarizes these results, demonstrating the effectiveness of StrokeFuse-AttnNet. The publicly available stroke dataset provides significant results for diagnostic practice. In addition, it demonstrates the model's capabilities and strengths in clinical settings.

Results on private dataset

To evaluate the effectiveness of the StrokeFuse-AttnNet model under practical conditions, the private dataset was used to determine the model's ability to distinguish between stroke and normal cases. Figure 9 shows the model's discriminative ability, achieving an area under the curve (AUC) of 0.9501. The model demonstrates a good balance of sensitivity and

Table 3 Hyperparameters and system configuration for training the proposed model

Hyperparameter	Value
Device	NVIDIA RTX 4070 GPU
Batch size	Dynamically adjusted (starting at 16)
Input image size	224 × 224 pixels
Normalization	Mean = [0.5], Standard deviation = [0.5]
Data augmentation	Rotation, flipping, cropping, brightness/contrast adjustments, random erasing
Learning rate	1e−4
Optimizer	Adam
Loss function	Binary cross-entropy
Scheduler	StepLR (Step size = 5 epochs, Gamma = 0.5)
Dropout rate	0.5
Epochs	100
Pretrained models	ResNet50, DenseNet121
Fine-tuned layers	Last 10 layers of each pretrained model
Attention mechanisms	Self-attention mechanism

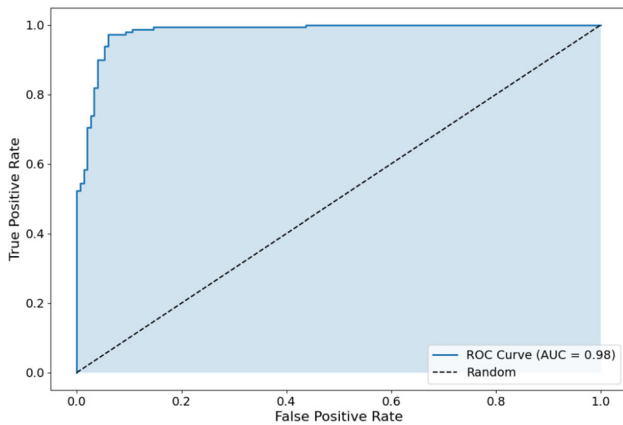


Fig. 7 Receiver operating characteristic (ROC) curve for the public dataset

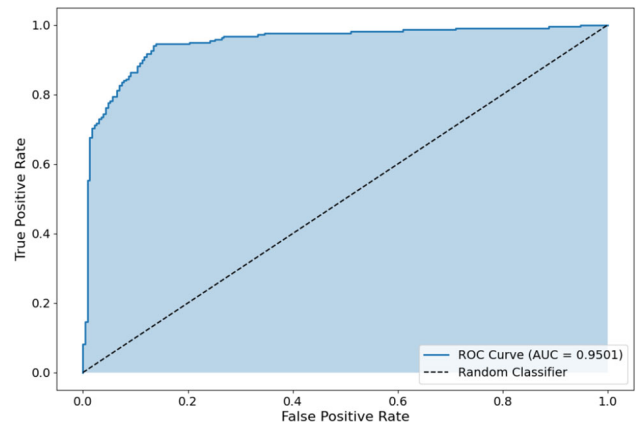


Fig. 9 Private dataset Receiver Operating Characteristic (ROC) curve

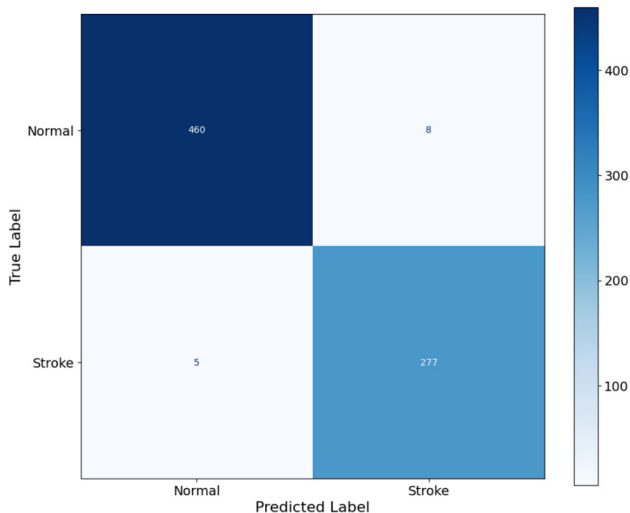


Fig. 8 The confusion matrix of the public data set is shown, along with information on true positives, false positives and false negatives

specificity, indicating its ability to reliably distinguish stroke cases from normal.

The model correctly classified 1255 normal samples, while 52 cases were incorrectly classified as strokes, as shown in Fig. 10. Among the stroke cases, 539 were correctly classified as strokes and 22 were incorrectly classified as normal. The model achieved an overall accuracy of 96.04% and the classification metrics confirm its robustness for normal cases: precision, recall and F1 score are 0.96, 0.96 and 0.96, respectively. For stroke cases, these values are 0.91, 0.96 and 0.93. Overall, the model achieves an average F1 score of 93.55% and a specificity of 96.02%, as shown in Table 6. Despite class imbalance, the model performed well on both metrics. The results indicate that StrokeFuse–AttnNet is generalizable and suitable for diagnostic use in a clinical setting, as demonstrated on private datasets.

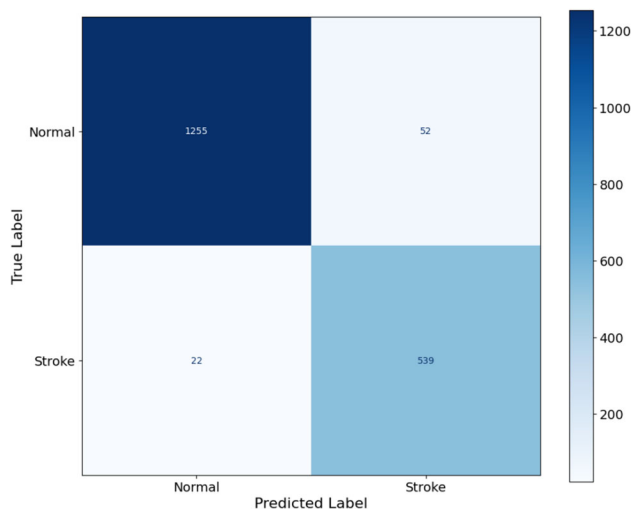


Fig. 10 Confusion matrix for the private dataset, detailing true positives, false positives and false negatives

Ablation study

The aim of an ablation study is to determine the contribution of various components and parameters which make up the StrokeFuse-AttnNet architecture, specifically self-attention, feature fusion and hyperparameter optimization. This study proposes to identify the essential components that contribute to stroke classification in CT images. The ablation analysis in this study follows a controlled one-factor-at-a-time strategy to evaluate the individual impact of key hyperparameters, including learning rate, batch size and dropout rate. This approach enables clear attribution of performance changes to specific parameters but does not explicitly model higher-order interactions between them. Conducting further systematic sensitivity analysis would improve performance and robustness of results, allowing optimisation of gradients through Bayesian or grid search. Incorporating such methods represents a promising direction for future work, particularly for deployment in heterogeneous clinical environments.

Effect of self-attention mechanism

The self-attention mechanism adjusts feature representations to increase the importance of the more relevant ones while reducing the importance of the less relevant ones. There was a significant drop in performance when the self-attention module was disabled. The performance in the public dataset dropped from 0.98 to 0.94, while in the private dataset it decreased from 0.9501 to 0.91. As the model becomes more capable of locating regions of importance in the image, the self-attention mechanism makes the features more discriminative and efficient.

Effect of feature fusion

The implementation of an augmented feature fusion technique that utilized feature maps by merging ResNet50 and DenseNet121 was key to recognizing the extensive variety of input features. The public AUC of ResNet50 was 0.93 and the private AUC was 0.89. However, the fused characteristics resulted in a significantly higher AUC of 0.98 for the public dataset and 0.9501 for the private dataset. The enhancement shows that combining feature representations from multiple backbones helps leverage the strengths of both.

Effect of hyperparameter optimization

Optimizing the main parameters led to improved and more stable performance. An experimentally determined learning rate of 1×10^{-4} proved optimal, representing a compromise between convergence speed and stability. Higher learning rates (1×10^{-3}) resulted in unstable training, while lower rates (1×10^{-5}) slowed convergence without significant performance improvement. A dropout rate of 0.5 led to overfitting, while rates of 0.3 and 0.7 resulted in underfitting and overfitting, respectively. The optimal batch size was 32, as smaller or larger sizes reduced performance. A batch size of 16 slowed training, while a size of 64 reduced accuracy and AUC.

The learning rate of 1×10^{-4} was selected as it provided stable convergence and consistently higher validation AUC compared to both higher (1×10^{-3}) and lower (1×10^{-5}) learning rates, as shown in the ablation results. A batch size of 32 provide optimal gradient stability, low computing cost and good generalization performance. Smaller batch sizes result in slower convergence, while larger batch sizes reduce overall accuracy. Similarly, a dropout rate of 0.5 effectively mitigated overfitting without under-regularizing the network, whereas lower or higher dropout values adversely affected validation performance. The results show that the selected hyper-parameters are a strong operating point for the proposed model.

Overall impact of the proposed framework

The robustness of the StrokeFuse-AttnNet architecture is demonstrated by the effective integration of the self-attention mechanism. The proposed model achieved an AUC of 0.98% for the public dataset and 0.9501 for the private dataset, demonstrating its generalization and stability. In addition, the public and private datasets were used to evaluate the model, achieving accuracy rates of 98.54% and 96.02%, respectively.

Table 4 Ablation study comparing ResNet50 and DenseNet121 using different configurations

Configuration	Public dataset (AUC)	Public dataset (accuracy)	Private dataset (AUC)	Private dataset (accuracy)
Proposed model (StrokeFuse-AttnNet)	0.98	98.54%	0.95	96.02%
Without self-attention	0.94	96.23%	0.91	94.10%
ResNet50	0.93	95.78%	0.89	93.50%
DenseNet121	0.91	94.89%	0.88	92.80%
Learning rate (1×10^{-3})	0.92	95.34%	0.90	93.89%
Learning rate (1×10^{-5})	0.95	97.12%	0.93	95.10%
Dropout rate (0.3)	0.94	96.45%	0.91	94.50%
Dropout rate (0.7)	0.93	95.67%	0.90	93.80%
Batch size (16)	0.95	97.10%	0.93	95.00%
Batch size (64)	0.92	95.12%	0.89	93.10%

Comparison with other deep learning methods

To evaluate and compare the performance of StrokeFuse-AttnNet with other deep learning models: ResNet50, DenseNet121 and EfficientNet-B0. We selected these models based on their efficacy in analysing medical images and their potential to serve as baseline architectures for comparison. All models were trained and tested with standardised pre-processing, data augmentation and evaluation metrics, using the same publicly available and private datasets. The results in Tables 5 and 6 show that StrokeFuse-AttnNet outperforms the baseline models. On the public dataset, StrokeFuse-AttnNet achieved an accuracy of 98.54% and an AUC of 0.98, demonstrating higher performance than ResNet50 (accuracy 95.78%, AUC 0.93), DenseNet121 (accuracy 94.89%, AUC 0.91) and EfficientNet-B0 (accuracy 93.80%, AUC 0.92). On the private dataset, StrokeFuse-AttnNet achieved an accuracy of 96.02% and an AUC of 0.9501, outperforming ResNet50 (accuracy 93.50%, AUC 0.89), DenseNet121 (accuracy 92.80%, AUC 0.88) and EfficientNet-B0 (accuracy 93.10%, AUC 0.89).

Based on the experimental results, StrokeFuse-AttnNet recognises various significant features in CT image data using fused hybrid features and self-attention mechanisms. The model generalises well and performs robustly on both datasets, suggesting its potential utility for stroke classification.

The use of traditional machine learning methods, such as support vector machines (SVMs) combined with hand-crafted or radiometric features, has been widely explored in stroke and medical image analysis alongside deep learning baselines. While these approaches provide advantages in interpretability and lower computational requirements, their performance depends heavily on feature engineering and they often struggle to capture complex spatial patterns and contextual information present in high-dimensional CT images. In contrast, the proposed end-to-end hybrid deep learning framework automatically learns hierarchical and complementary feature representations, enabling superior discrimination performance and robustness across heterogeneous datasets.

Comparison of strokefuse-attnnet with the state-of-the-art (SOTA)

The state-of-the-art (SOTA) results summarized in Table 7 are reported from previously published studies and were obtained using different datasets, imaging protocols and preprocessing pipelines. Therefore, these comparisons are intended to provide contextual benchmarking rather than a direct, head-to-head performance evaluation. Direct comparisons under identical experimental settings are provided separately for baseline deep learning models (ResNet50,

DenseNet121 and EfficientNet-B0), which were re-trained and evaluated on the same datasets using identical pre-processing and evaluation protocols. The results of the comparative analysis are shown in Table 7. Due to the efficient architecture of StrokeFuse-AttnNet, other competing methods for stroke detection cannot achieve improved stroke detection. These findings demonstrate the advantages of hybrid feature fusion and self-attention for accurate stroke detection in CT images, indicating that these models are promising for clinical decision support in stroke treatment. Unlike other models, StrokeFuse-AttnNet uses these ResNet50 and DenseNet121 architectures. The selected architecture captures the necessary global and local features to process the required medical image. StrokeFuse-AttnNet effectively harnesses the distinct feature representations from ResNet50 and DenseNet121. The synthetic additive fusion mechanism provides high versatility in creating enhanced composite representations. This integrated method may more accurately classify strokes and non-strokes.

Moreover, adopting a self-attention mechanism can enhance the classification ability of the fused features by focusing on relevant spatial and channel information while ignoring unnecessary information. By leveraging the finer patterns and phenomena identified during classification, the model can predict various outcomes with greater sensitivity and specificity. Figure 11 shows that StrokeFuse-AttnNet is the most efficient. With fewer numerical parameters, reduced floating point operations and shorter test times, this model performs well. The proposed model for stroke diagnosis in CT scans enhances medical image analysis. It is robust and effective.

The high performance with a few parameters is the key strength of StrokeFuse-AttnNet. A hybrid feature fusion strategy and attention mechanism are used to limit overfitting and help retain generalization performance on unseen and new data. StrokeFuse-AttnNet, capable of detecting stroke, is a reliable real-time detection model that is both computationally efficient and accurate.

Discussion

The proposed stroke detection model was evaluated using public and private datasets to assess its predictive performance. Testing was conducted using a consistency protocol and a holdout strategy, with 70% of the data allocated for training and 30% for testing. Model training was conducted exclusively on the training set, while all reported results were obtained from the independent test set. On the public dataset, the proposed StrokeFuse-AttnNet achieved an accuracy of 98.27% and an AUC of 0.98, outperforming existing approaches evaluated on the same dataset. The model demonstrated strong generalization capability by correctly classifying 98.3% of normal cases and 98.2% of stroke

Table 5 Comparison of results between the proposed and other deep neural networks on a public dataset

Model	Accuracy	AUC	F1-score	Specificity
ResNet50	95.78%	0.93	95.20%	96.30%
DenseNet121	94.89%	0.91	94.10%	95.20%
EfficientNet-B0	93.80%	0.92	93.40%	94.50%
StrokeFuse-AttnNet (proposed)	98.27%	0.98	97.70%	98.29%

Table 6 Comparison of results between the proposed and other deep networks on a private dataset

Model	Accuracy	AUC	F1-score	Specificity
ResNet50	93.50%	0.89	92.80%	94.00%
DenseNet121	92.80%	0.88	91.90%	93.10%
EfficientNet-B0	93.10%	0.89	92.40%	93.60%
StrokeFuse-AttnNet (proposed)	96.04%	0.95	93.55%	96.02%

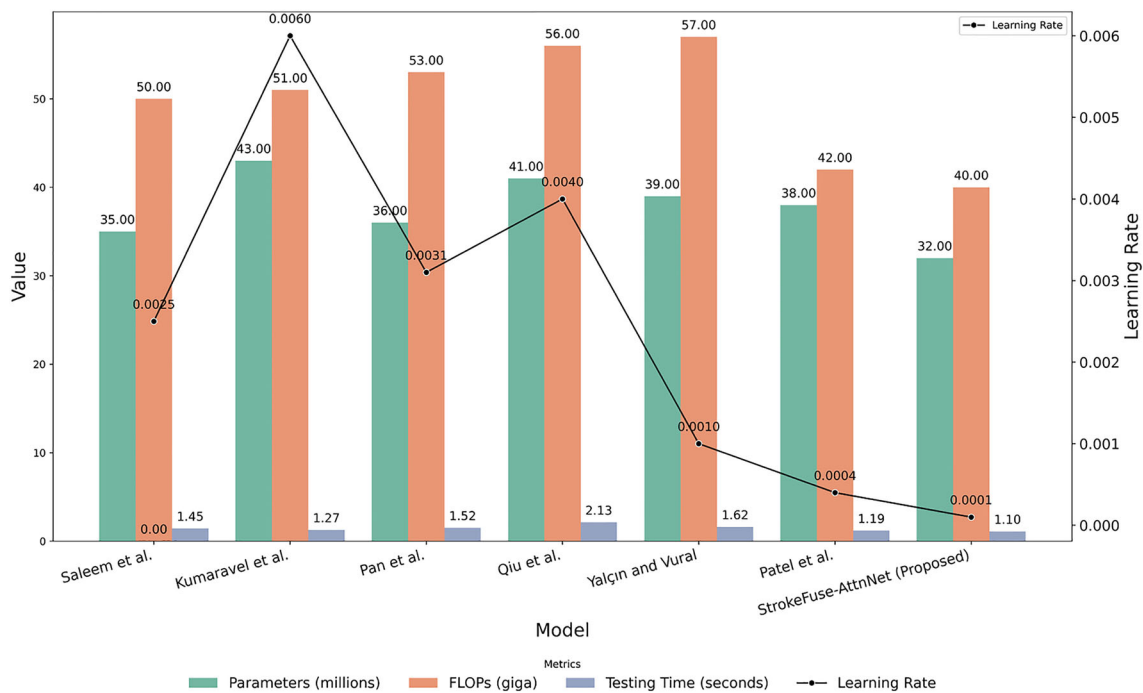


Fig. 11 The complexity of the various models in the computations supports the effectiveness of the proposed model

cases. In the private dataset, we achieve an accuracy of 96.04%, an AUC of 0.9501 and robust class-wise performance, despite increased clinical heterogeneity. The results shows that the proposed framework scales well and is useful for diagnosis on real data.

StrokeFuse-AttnNet is designed as a computationally efficient model suitable for real-time clinical use. The model has 32 million trainable parameters and a per-inference cost of about 40 GFLOPs, resulting in a modest memory footprint suitable for modern workstation GPUs, which are typically available in hospitals. On the experimental hardware used in this study (NVIDIA RTX 4070 GPU), the average inference latency was tens of milliseconds per CT image, supporting high-throughput processing in routine

clinical workflows. Due to its end-to-end architecture and reliance solely on standard non-contrast CT images, the proposed model can function as a backend clinical decision support module and can be incorporated into existing PACS. This enables automated stroke screening to be performed transparently alongside routine radiological review without disrupting established diagnostic procedures. As with most learning-based imaging systems, performance may degrade in the presence of severe motion artefacts, scans of very low resolution, or unusual acquisition protocols. These possible failure cases highlight the necessity for quality checks and future domain-adaptive fine-tuning to ensure robust deployment across diverse clinical settings.

Table 7 Comparative analysis results of the (SOTA) model on the public dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)
Pan et al. [20]	97.7	97.2	97.6	97.4	98.1
Saleem et al. [14]	95.5	92.0	90.8	95.0	–
Abdi et al. [47]	97.2	97.0	97.0	97.0	–
Qiu et al. [48]	97.9	96.5	96.8	96.6	98.3
Saleem et al. [2]	97.8	98.5	–	–	–
Irfan et al. [4]	95.8	95.8	–	95.7	–
Yalçın and Vural [49]	98.3	97.1	97.6	97.3	96.5
Kumaravel et al. [50]	98.6	97.0	97.2	97.1	97.1
Patel et al. [22]	97.7	96.5	96.7	96.6	98.0
Saleem et al. [14]	96.5	98.0	93.5	96.0	–
StrokeFuse-AttnNet (Proposed)	98.27	97.19	98.23	97.70	98.29

The performance of StrokeFuse-AttnNet is due to its hybrid architectural design, which combines global semantic representations from ResNet50 with fine-grained feature reuse from DenseNet121, further refined by a self-attention mechanism that highlights diagnostically relevant spatial and channel information. Compared to single-architecture models, the hybrid approach provides a richer feature representation and is more computationally efficient, allowing deployment in both high- and low-resource clinical settings. Despite its effectiveness, certain failure cases may occur in real-world deployment, such as CT scans affected by patient motion artefacts, low spatial resolution, severe noise, or atypical anatomical presentations. These factors can degrade feature quality and reduce classification confidence. To address such challenges, the proposed framework can be adapted through enhanced preprocessing (such as noise suppression and resolution normalization), targeted data augmentation and fine-tuning on institution-specific datasets. In the future, we will work on extending multi-modal incorporation of MRI data and multi-class predictions of stroke subtypes.

Interpretability was assessed using the Grad-CAM and saliency visualizations shown in Fig. 6. Although self-attention maps are not displayed separately, this will reflect the attention-refined activation regions. In stroke-positive cases, these regions generally correspond to clinically relevant patterns such as infarct zones, haemorrhagic areas and hemispheric asymmetries, whereas normal cases show diffuse, low-intensity activations. This qualitative behavior suggests that the self-attention mechanism directs the model's focus towards diagnostically relevant areas, which supports the interpretability of the model, although this has not been quantitatively verified.

In stroke-positive CT images, the highlighted regions consistently correspond to clinically relevant patterns such as hypo-dense infarct areas, haemorrhagic regions and asym-

metries typically associated with acute stroke. In contrast, normal cases exhibit diffuse or minimal activation, indicating the absence of localized pathological features. This behaviour suggests that the self-attention mechanism effectively directs the network to diagnostically meaningful areas. The attention maps not only provide visual confirmation of the model but also highlight diagnostic features that correspond with radiological signs of stroke. For example, in the case of ischaemic stroke, we consistently observed emphasis on low-attenuation regions of the MCA. Negative cases show diffuse or no attention, consistent with the model's response to non-lesion inputs. This is important for clinical use, as visual and spatial interpretation increases user awareness of the automated model. Furthermore, attention maps help identify potential false positives and improve the model for cases with ambiguous results near the decision boundary. Although this study addresses binary stroke detection, the proposed StrokeFuse-AttnNet can be directly extended to multi-class stroke subtyping (e.g., ischaemic vs. haemorrhagic) by replacing the final sigmoid layer with a softmax classifier, while retaining the hybrid fusion and self-attention mechanisms.

Table 4 shows that although StrokeFuse-AttnNet achieves higher classification accuracy than generic models, it also addresses key limitations of previous approaches. Using multiple CNNs with a single architecture, such as VGG16 or EfficientNet-B0, results in overfitting and poor generalization when applied to heterogeneous medical images. Furthermore, conventional deep models often lack spatial interpretability mechanisms and cannot be used clinically, as explainable models are essential in this context. The proposed approach combines the learning of diverse visual attribute properties through two networks. ResNet50 is a widely used neural network architecture for designing deep neural networks that have been previously trained. ResNet50 extracts rich global features from input images. Its primary

focus is the design of efficient residual blocks to address the issue of vanishing gradients. ResNet50 has a lower computational cost but remains powerful. It does not change the dimension of the input, only adding depth. By combining attention-guided feature refinement with clinically interpretable visualization techniques, the proposed framework enhances transparency and supports trust in automated stroke detection for real-world clinical deployment. The inference latency reflects performance on a clinical workstation equipped with a GPU. In future work, we will benchmark on edge devices and directly compare with lightweight architectures such as MobileNetV3 or EfficientNet-B0 under the same conditions. The absence of detailed demographic and clinical metadata, such as stroke subtype and onset timing, is a limitation of the present study and will be addressed in future work through richer multi-modal clinical data integration.

Conclusion

This study proposed a novel StrokeFuse-AttnNet for stroke detection using CT images. The design employed a feature fusion approach combining ResNet50 and DenseNet121. The self-attention mechanism improves the model by highlighting relevant features and reducing the impact of irrelevant ones. StrokeFuse-AttnNet performs equally well on both public and private datasets. The model achieves an accuracy of 98.27% and an AUC of 0.98 on the public dataset and an accuracy of 96.04% with an AUC of 0.9501 on the private dataset. This model performs better than the state of the art. With 32 million parameters and 40 GigaFLOPs, it runs efficiently for real-time clinical applications. Its consistency, predictability and balanced computational complexity provide an effective solution to major challenges in medical imaging. In future studies, we will incorporate additional imaging techniques alongside current ones (such as magnetic resonance imaging) and extend our model's predictive accuracy to include more stroke classes. Consequently, the model is likely to become more clinically useful. StrokeFuse-AttnNet can improve stroke detection, clinical decision-making and overall patient outcomes.

Acknowledgements This research has received funding support from the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [Grant No. B04G640068]. This research is also funded by Thailand Science Research and Innovation Fund (HEA_FF_69_103_2100_016 and IND_FF_69_196_2100_033) and Ratchadapisek Somphot Fund for the Center of Excellence in Artificial Intelligence, Machine Learning and Smart Grid Technology and by the Second Century Fund (C2F), Chulalongkorn University.

Data Availability The data supporting the results of this study are available on request from the corresponding author. The source code and

model implementation for this study are publicly available at: <https://github.com/asim-lab/Hybrid-ResNet-DenseNet-Stroke-CT>.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval This study used retrospectively collected, fully anonymized non-contrast CT images that were originally acquired during routine clinical care. No identifiable patient information, clinical records, or metadata were accessible to the authors. The study did not involve prospective data collection, patient interaction, or clinical intervention and all analyses were conducted in accordance with institutional data protection policies and applicable ethical guidelines.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Akter B et al (2022) A machine learning approach to detect the brain stroke disease. In: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), organization IEEE, pp 897–901
2. Saleem MA et al (2025) Enhancing stroke risk prediction through class balancing and data augmentation with cbda-resnet50. *Sci Rep* 15:24553
3. Ozaltin O, Coskun O, Yeniay O, Subasi A (2022) A deep learning approach for detecting stroke from brain ct images using oznet. *Bioengineering* 9:783
4. Irfan M, Subasi A, Mustafa N, Westerlund T, Chen W (2024) An evaluation of pretrained convolutional neural networks for stroke classification from brain ct images. In: Applications of artificial intelligence in healthcare and biomedicine. Elsevier, pp 111–135
5. Saleem MA et al (2024) An intelligent learning system based on electronic health records for unbiased stroke prediction. *Sci Rep* 14:23052
6. Shen D, Wu G, Suk H-I (2017) Deep learning in medical image analysis. *Annu Rev Biomed Eng* 19:221–248
7. Litjens G et al (2017) A survey on deep learning in medical image analysis. *Med Image Anal* 42:60–88
8. Chawla M, Sharma S, Sivaswamy J, Kishore LA (2009) method for automatic detection and classification of stroke from brain ct images. In: 2009 Annual international conference of the IEEE engineering in medicine and biology society. IEEE, pp 3581–3584

9. Korra S, Soora N, Jahan T, Ramana N, Rajesh A (2024) Brain ct image processing using u-net model with data augmentation for detection of ischemic and haemorrhage strokes. *Int J Intell Syst Appl Eng* 12:72–82
10. Sailasya G, Kumari GLA (2021) Analyzing the performance of stroke prediction using ml classification algorithms. *Int J Adv Comput Sci Appl* 12
11. Cortés-Ferre L, Gutiérrez-Naranjo MA, Egea-Guerrero JJ, Pérez-Sánchez S, Balcerzyk M (2023) Deep learning applied to intracranial hemorrhage detection. *J Imaging* 9:37
12. Yeo M et al (2023) Evaluation of techniques to improve a deep learning algorithm for the automatic detection of intracranial haemorrhage on ct head imaging. *Eur Radiol Exp* 7:17
13. Tursynova A et al (2023) Deep learning-enabled brain stroke classification on computed tomography images. *Comput Mater Contin* 75:1431–1446
14. Saleem MA et al (2024) Innovations in stroke identification: a machine learning-based diagnostic model using neuroimages. *IEEE Access*
15. Jin B, Xu X (2025) Employing gaussian process regression with Bayesian inference to predict the living-materials producer price index in China. *Qual Quant*:1–39
16. Jin B, Xu X (2025) Predicting the trading volume of the thermal coal futures through gaussian process regressions. *J Uncertain Syst*:2550019
17. Xu X, Zhang Y (2022) Thermal coal price forecasting via the neural network. *Intell Syst Appl* 14:200084
18. Xu X, Zhang Y (2021) Network analysis of corn cash price comovements. *Mach Learn Appl* 6:100140
19. Xu X, Zhang Y (2023) A high-frequency trading volume prediction model using neural networks. *Decis Anal J* 7:100235
20. Pan J et al (2021) Detecting the early infarct core on non-contrast ct images with a deep learning residual network. *J Stroke Cerebrovasc Dis* 30:105752
21. Chen Y-T et al (2022) Deep learning-based brain computed tomography image classification with hyperparameter optimization through transfer learning for stroke. *Diagnostics* 12:807
22. Patel CH, Undaviya D, Dave H, Degadwala S, Vyas D (2023) Efficientnetb0 for brain stroke classification on computed tomography scan. In: 2023 2nd International conference on applied artificial intelligence and computing (ICAIC). IEEE, pp 713–718
23. Çinar N, Kaya B, Kaya M (2023) Brain stroke detection from ct images using transfer learning method. In: 2023 13th International conference on advanced computer information technologies (ACIT). IEEE, pp 595–599
24. Abumihsan A et al (2025) A novel hybrid model for brain ischemic stroke detection using feature fusion and convolutional block attention module. *IEEE Access*
25. Zhou C, Feng D, Chen S, Ban N, Pan J (2024) Portable vision-based gait assessment for post-stroke rehabilitation using an attention-based lightweight cnn. *Expert Syst Appl* 238:122074
26. Aksoy S, Demircioglu P, Bogrekcı I (2024) Optimizing stroke classification with pre-trained deep learning models. *J Vasc Dis* 3:480–494
27. Vindas Y, Guépié BK, Almar M, Roux E, Delachartre P (2022) An hybrid cnn-transformer model based on multi-feature extraction and attention fusion mechanism for cerebral emboli classification. In: Machine learning for healthcare conference. PMLR, pp 270–296
28. Abulfaraj AW, Dutta AK, Sait ARW (2024) Feature fusion-based brain stroke identification model using computed tomography images. *J Disabil Res* 3:20240060
29. Jin B, Xu X (2025) China commodity price index (ccpi) forecasting via the neural network. *Int J Financ Eng*:1–27
30. Jin B, Xu X (2025) High-frequency csi300 spot and futures price predictions via the neural network. *J Uncertain Syst*:2550008
31. Xu X, Zhang Y (2021) Corn cash price forecasting with neural networks. *Comput Electron Agric* 184:106120
32. Jin B, Xu X (2025) Forecasts of coking coal futures price indices through gaussian process regressions. *Miner Econ* 38:203–217
33. Jin B, Xu X (2024) Machine learning coffee price predictions. *J Uncertain Syst* 17:2450023
34. Xu X, Zhang Y (2023) Price forecasts of ten steel products using gaussian process regressions. *Eng Appl Artif Intell* 126:106870
35. Jin B, Xu X (2025) A study of contemporaneous residential real estate price causation across major Jiangsu province cities: methodology using vector error-correction models and directed acyclic graphs. *Econ Open*:2550008
36. Xu X, Zhang Y (2023) An integrated vector error correction and directed acyclic graph method for investigating contemporaneous causalities. *Decis Anal J* 7:100229
37. Xu X (2019) Contemporaneous and granger causality among us corn cash and futures prices. *Eur Rev Agric Econ* 46:663–695
38. Xu X (2020) Corn cash price forecasting. *Am J Agr Econ* 102:1297–1320
39. Xu X, Zhang Y (2021) Individual time series and composite forecasting of the Chinese stock index. *Mach Learn Appl* 5:100035
40. Xu X (2017) Short-run price forecast performance of individual and composite models for 496 corn cash markets. *J Appl Stat* 44:2593–2620
41. Mohammed FA, Tune KK, Assefa BG, Jett M, Muhie S (2024) Medical image classifications using convolutional neural networks: a survey of current methods and statistical modeling of the literature. *Mach Learn Knowl Extr* 6:699–735
42. Alwakid G, Tahir S, Humayun M, Gouda W (2024) Improving Alzheimer's detection with deep learning and image processing techniques. *IEEE Access*
43. Rahman A (2022) Brain stroke ct image dataset. <https://www.kaggle.com/datasets/afriDIRahman/brain-stroke-ct-image-dataset>. Accessed 23 Oct 2024
44. Hassan SM, Maji AK (2024) Pest identification based on fusion of self-attention with resnet. *IEEE Access* 12:6036–6050
45. Xu X, Zhang Y (2023) A gaussian process regression machine learning model for forecasting retail property prices with Bayesian optimizations and cross-validation. *Decis Anal J* 8:100267
46. Jin B, Xu X (2025) Machine learning-based forecasts of residential property prices in Hangzhou city, Zhejiang province, china. *Neural Comput Appl* 37:4971–4988
47. Abdi H, Sattar MU, Hasan R, Dattana V, Mahmood S (2025) Stroke detection in brain ct images using convolutional neural networks: model development, optimization and interpretability. *Information* 16:345
48. Qiu W et al (2020) Machine learning for detecting early infarction in acute stroke with non-contrast-enhanced ct. *Radiology* 294:638–644
49. Yalçın S, Vural H (2022) Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks. *Comput Biol Med* 149:105941
50. Kumaravel P, Mohan S, Arivudaiyanambi J, Shajil N, Venkatakrishnan HN (2021) A simplified framework for the detection of intracranial hemorrhage in ct brain images using deep learning. *Curr Med Imaging* 17:1226–1236

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.