StrokeNetBench: A Comparative Framework of Deep Architectures for Stroke Detection and Classification

Mashuka Bashar Chowdhury, and Sadia Jannat Mitu*

Department of Computing Science and Engineering, Daffodil International University, Dhaka-1216, Bangladesh

chowdhury15-5146@diu.edu.bd, jannatmitu.cse@diu.edu.bd*

Abstract. Stroke is a leading cause of mortality and permanent disability worldwide. It is a serious neurological emergency. Timely intervention and better outcomes depend on accurate identification of its two major types, ischemic and hemorrhagic stroke. Manual interpretation of brain computed tomography (CT) scans is laborious, subjective, and error-prone. We provide a deep learning-based architecture for automatically classifying brain CT scans into three groups: ischemia, bleeding (hemorrhage), and no stroke (control). Our approach automatically constructs features from CT scans using convolutional neural networks (CNNs) and transfer learning. The framework incorporates preprocessing, data augmentation, and class imbalance handling using class weighting and focal loss. A custom CNN, VGG16, VGG19, ResNet50 and ViT-Base are evaluated under various layer configurations. Models are assessed using 10-fold cross-validation. ResNet50 with a single unfrozen residual block achieves the best performance, with 97.3% accuracy, 95.7% precision, 96.4% recall, 96.1% F1-score, and low false negative rates. High class separability is shown by ROC-AUC scores of 99% (No Stroke), 91% (Bleeding), and 97% (Ischemia). Grad-CAM-based visual explanations confirm the model focuses on stroke-relevant areas, enhancing interpretability. These results demonstrate deep learning models' effectiveness for reliable stroke diagnosis, supporting clinical decision-making.

Keywords: Stroke Detection, CT Imaging, Deep Learning, VGG16, VGG19, ResNet50, ViT-Base, Focal Loss, Class Weighting, Transfer Learning, Convolutional Neural Network, 10-Fold Cross-Validation, Grad-CAM.

1 Introduction

A stroke is a sudden, serious medical disease that occurs when blood flow to the brain is interrupted. If treatment is delayed, it can cause irreversible neurological impairment or even death [1]. With an estimated 15 million stroke victims and about 5 million fatalities each year, it is the second greatest cause of death worldwide and one of the main causes of long-term impairment [2]. Strokes place a significant economical burden on victims due to the requirement for long-term

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_10

care. Clinically, they are primarily divided into two subtypes: hemorrhagic and ischemic [3].

About 85% of strokes are ischemic strokes, which happen when a blood clot stops an artery supplying the brain, depriving the neural tissue of oxygen and nutrition [4]. On the other hand, hemorrhagic stroke happens when weak blood arteries in the brain burst, resulting in bleeding that raises intracranial pressure and damages the brain [5]. Hemorrhagic and ischemic strokes can have similar neurological symptoms, yet they have quite different underlying causes and treatment. It is crucial to differentiate quickly and accurately since thrombolytics, which are useful for ischemic stroke, might increase the risk of bleeding in hemorrhagic instances [6].

Both CT and MRI are necessary for the diagnosis of stroke subtypes, and because of their accessibility and speed, CT scans are frequently performed first in acute instances [7]. Diagnostic delays are frequently caused by the subjective and time-consuming nature of manual CT interpretation, particularly in situations with a large patient load or few specialists [8]. Deep learning (DL), one of the most recent developments in artificial intelligence, offers potential ways to automate medical image processing [9]. Without the need for human feature extraction, deep learning models like CNNs can automatically identify complex patterns from medical images, including those for cancer and neurological illnesses. These models frequently match or even outperform skilled radiologists in this regard [10]. DL can facilitate rapid, objective, and accurate CT image interpretation for stroke diagnosis, supporting clinical decision-making. However, model development and generalization remain challenging due to small and imbalanced datasets, particularly the limited number of hemorrhagic stroke cases and variability in imaging conditions [11].

The majority of current studies, despite recent advancements, concentrate on the binary classification of stroke subtypes, frequently disregarding the inclusion of control cases (no stroke). Additionally, there is a lack of systematic comparisons between transfer learning models and custom CNNs that employ rigorous validation strategies. Furthermore, the underrepresentation of hemorrhagic stroke remains a significant challenge that has not been adequately addressed in numerous prior studies, as well as imbalanced datasets. Moreover, there are only a few studies that have utilized Grad-CAM for classifying strokes, resulting in a lack of explainable AI methods in this area.

This paper's primary contributions include the development of a multiclass CT classification framework that can differentiate between cases of ischemic stroke, hemorrhagic stroke, and no stroke. We systematically evaluate how well a bespoke CNN performs in comparison to popular model like VGG16, VGG19, ResNet50, and ViT-Base under various fine-tuning strategies. Both class weighting and focal loss are incorporated into the framework to mitigate class imbalance. An independent test set and stratified 10-fold cross-validation are used to thoroughly assess the proposed methods. Based on all evaluation metrics, our results demonstrate that ResNet50 with partial fine-tuning consistently performs better. To improve interpretability, Grad-CAM (Gradient-weighted Class Acti-

vation Mapping) is used to visualize and highlight important regions in the CT images, showing that the model focused on areas affected by stroke.

This paper's remaining sections are organized as follows: Related research on stroke detection is covered in Section 2. An description of the suggested methodology is provided in Section 3. Performance simulation is displayed in Section 4. The research's conclusions are presented in Section 5.

2 Literature Review

Stroke is a major global cause of death and permanent disability, and successful treatment depends on an early and precise diagnosis. In recent years, there has been an increasing amount of research on automated stroke recognition and subtype categorization using machine learning and deep learning algorithms applied to medical imaging such as CT and MRI scans.

Traditional machine learning techniques were the mainstay of early automated stroke detection research[12]. UmaMaheswaran et al. [13] showed that, More sophisticated models with optimal feature selection, such XGBoost, obtained noticeably higher accuracy than conventional classifiers like SVM, Random Forest, and shallow ANNs trained on manually created features from CT scans, which only exhibited slight performance gains. however, their reliance on manually created descriptors hindered their ability to generalize their findings across datasets. These findings demonstrated how shallow feature representations are insufficient to capture the complexity of medical imaging data.

Convolutional neural networks (CNNs) have shown an improved ability to learn abstract features directly from raw images as deep learning has become more popular. While Chauhan et al. [14] pointed out that shallow CNNs continued to perform poorly when compared to deeper designs, especially in ischemic stroke detection. Neethi et al. [11] developed a CNN that outperformed conventional methods. This prompted the classification of strokes using deeper networks and more complex architectures.

Sahoo et al. [15] demonstrated that, pretrained models like VGG16 and ResNet50 to increase accuracy and resilience despite the lack of medical data. Dataset imbalance, however, continued to be a major obstacle. Merdas et al. [16] presented the EMS model, in order to enhance stroke risk prediction on unbalanced datasets, which combines Elastic Net, MLP, and SMOTE. SMOTE improved model balance and performance, laying a solid basis for additional study to improve patient warning systems and forecast accuracy whereas Lin et al. [17] developed focal loss to improve the detection of uncommon stroke forms, such as hemorrhagic stroke, by concentrating training on minority classes that are difficult to identify. This helps to overcome class imbalance in stroke datasets. Mundargi et al. [18] highlighted the use of k-fold cross-validation to lessen overfitting and produce more accurate results for stroke classification.

More recent studies have explored advanced architectures and strategies. Arbabshirani et al. [19] presented a 3D convolutional neural network that can automatically identify cerebral hemorrhage on head CT scans. The model may

be integrated into hospital workflows to enable quick triage. Eshmawi et al. [20] suggested a CAD-BSDC model for MRI-based stroke classification, using an ensemble of CNN feature extractors with metaheuristic optimization, outperforming current techniques in terms of accuracy and efficiency.

Even with these developments, a number of problems still exist. Imbalanced datasets, inadequate validation, or a restricted focus on binary classification limit the scope of many studies. Furthermore, for multiclass stroke classification (no-stroke, ischemia, bleeding), there are notably few systematic comparisons between custom CNNs and contemporary transfer learning models, such as transformer-based architectures. Moreover, there are only a few studies that have utilized Grad-CAM for classifying strokes. Additionally, existing studies typically focus on a single metric, like ROC-AUC, instead of reporting a whole set of evaluation metrics (accuracy, precision, recall, F1-score, false negative rate), which are essential for clinically meaningful assessment.

3 Methodology

Extensive research has been conducted on the use of deep learning in conjunction with CT imaging for diagnosing and classifying strokes. Our method incorporates tailored preprocessing, effective data augmentations, class-imbalance handling (Focal Loss with class weighting), careful dataset selection, and systematic model training and evaluation. Figure 1 showcases this detailed pipeline, emphasizing the systematic framework that ensures the precision and dependability of our classification system.

3.1 Dataset Collection

The dataset utilized in this research is sourced from a publicly available Kaggle repository [21], initially compiled by the Republic of Türkiye Ministry of Health and meticulously annotated by seven radiologists for the TEKNOFEST-2021 competition [22]. It consists of 6,650 anonymized grayscale brain CT slices divided into three categories: No Stroke (4,427), Ischemia (1,130), and Bleeding (1,093). While images in the No Stroke category are provided solely in DICOM and PNG formats, the Ischemia and Bleeding categories also include overlay files that accurately mark the affected areas. The dataset does not include any patient demographic information. For this study, only the PNG images are used, systematically organized into folders based on class, and visually checked to ensure the accuracy and clarity of the labels. The external test set is not included as it contains binary labels that do not align with the study's focus on multi-class classification.

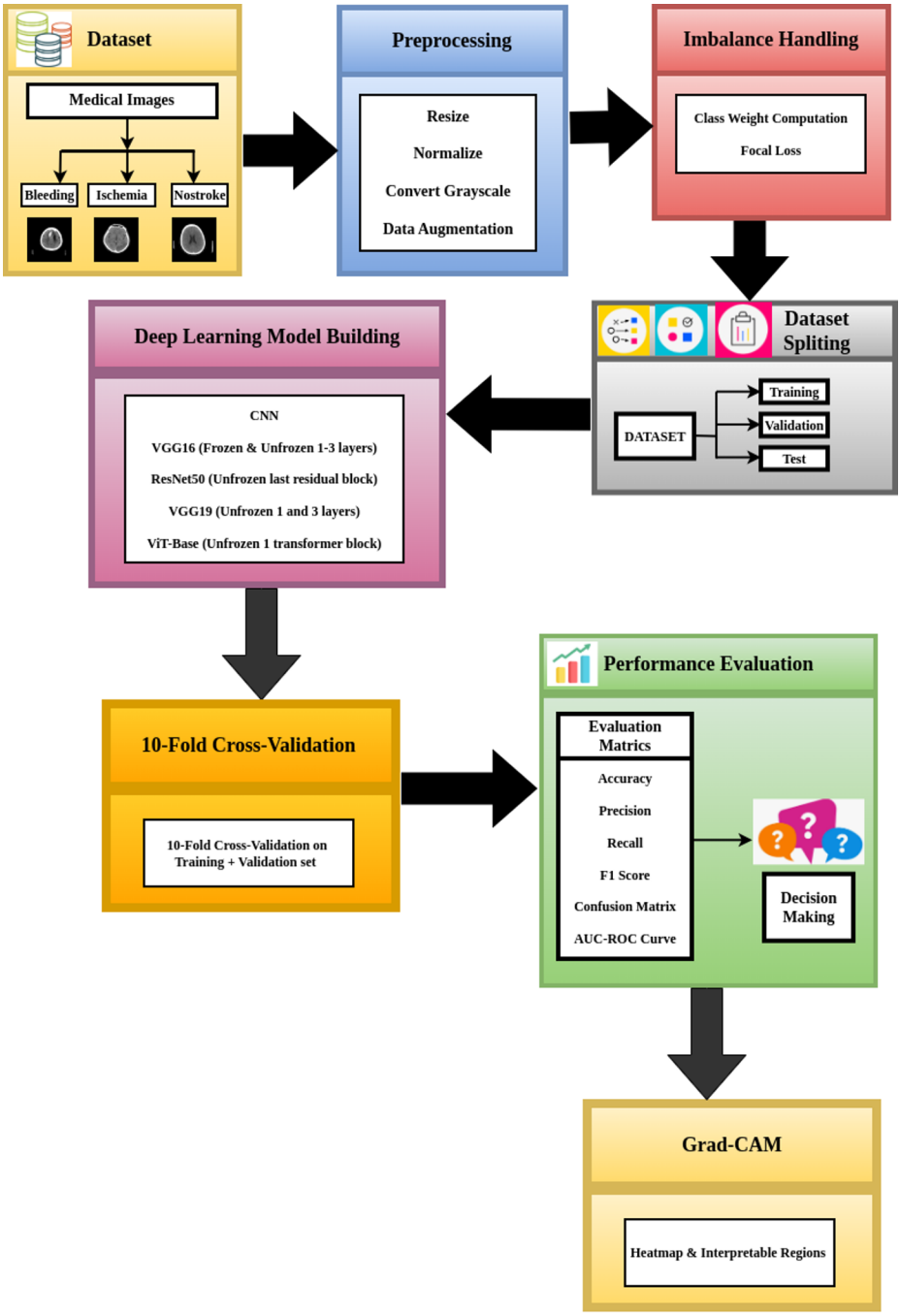


Fig. 1: Work Flows for Proposed Methodology.

3.2 Data Preprocessing

The CT scans are extracted from PNG format and adjusted to a size of 224×224 pixels. As the images are in grayscale, the single channel is replicated to create three-channel RGB images, making them compatible with pretrained networks. Pixel values are scaled to a range of [0, 1] by dividing by 255 in order to improve training stability. A fixed crop of 60 pixels from each side is performed to eliminate unnecessary borders and concentrate on the brain areas. Lastly, class labels are converted to one-hot encoding for multiclass classification.

3.3 Data Augmentation

To improve generalization and minimize overfitting while preserving anatomical precision, we implement a set of carefully chosen augmentations. These include the addition of Gaussian noise, adjustments to brightness and contrast, slight rotations of up to ±5°, minor translations, and zooming within a ±10% range. We deliberately avoid more drastic transformations, such as flipping and large rotations, to ensure the structural realism remained intact.

3.4 Feature Engineering

Using preprocessed CT scans, deep learning algorithms automatically extract relevant features, removing the need for human feature extraction. This method takes advantage of the models' capability to discern intricate patterns crucial for precise stroke classification.

3.5 Imbalance Handling

An in-depth exploratory data analysis (EDA) of the dataset uncover a significant class imbalance of 4:1:1, with No Stroke (4,427), Bleeding (1,093), and Ischemia (1,130) among 6,650 samples. This imbalance poses a risk of skewing the model towards the predominant class. To mitigate this, we apply class weights that are inversely related to class frequencies and utilized focal loss to emphasize the minority samples that are challenging to classify. The focal loss parameter (α) is meticulously adjusted to work in harmony with the class weights, promoting balanced learning and minimizing overfitting.

3.6 Dataset Splitting

The dataset is split into training, validation, and test sets with different ratios—65:20:15, 75:15:10, and 70:15:15—to ensure an even distribution of classes. The 70:15:15 division consistently yield the best outcomes across all metrics—accuracy, precision, recall, and F1-score. It is important to note that the test set is kept apart and utilize exclusively for final evaluation.

3.7 Deep Learning Model Building

Using a custom CNN and transfer learning with VGG16, VGG19, ResNet50, and ViT-Base, we develope a stroke classification system. Each of the three convolutional layers (32, 64, and 128 filters) in the custom CNN followed by max-pooling, a dense ReLU layer with 256 units and dropout 0.5, and a softmax output. We replace the top layers of VGG16 and VGG19, which are pretrained on ImageNet, then gradually unfrozen the final 1-3 convolutional layers. All other layers stay frozen, however the last transformer block of ViT-Base and the final residual block of ResNet50 are fine-tuned. Global average pooling, a dense ReLU layer (256 units), dropout 0.5, and softmax classification are applied to each transfer model before it is finalized.

3.8 Training Setup

The Adam optimizer is used to train the models with a batch size of eight and a learning rate of 1e-4 in order to ensure effective and stable convergence. The ideal weights based on validation loss are restored by applying early stopping with five epoch patience to prevent overfitting. More accurate training is also made possible by using a ReduceLROnPlateau scheduler to lower the learning rate when the validation loss stay constant for three consecutive epochs.

All architectures are trained using the same preprocessing, batch size, optimizer, learning rate schedule, and evaluation metrics to ensure a systematic comparison between the transfer learning and bespoke CNN models.

3.9 10 Fold Cross Validation

Models are trained on nine of the folds and verified on the remaining one in each cycle using a stratified 10-fold cross-validation technique in the training-validation set. Class weights are established for each training fold in order to address class imbalance. To assess the overall robustness of the model, metrics including accuracy, precision, recall, and F1-score are computed for each fold and then averaged. To ensure an objective assessment of the models' generalizability, the independent test set is reserved solely for the final performance evaluation.

3.10 Grad-CAM for Model Interpretability

To enhance the interpretability of the deep learning models, Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to the last convolutional layer of the best-performing model, ResNet50. Grad-CAM generates heatmaps that highlight the regions in CT images that the model considers most important for stroke classification. By overlaying these heatmaps on the original images, we visually assessed whether the model focused on clinically relevant stroke-affected areas.

Grad-CAM visualizations are produced on the validation set during 10-fold cross-validation to examine model attention across different data splits. Additionally, the saved best model is used to generate Grad-CAM on the held-out test

set, providing interpretable results on unseen data and supporting the clinical trustworthiness of the automated predictions.

4 Result Analysis

In this paper, we use stratified 10-fold cross-validation to evaluate the performance of our deep learning models for multiclass stroke classification. Several metrics, such as accuracy, precision, recall, F1-score (weighted and macro), and false negative rate (FNR), are calculated to assess the model's overall and class-wise efficacy. Performance findings are compiled into tables and displayed using ROC curves, confusion matrices, and comparison charts to provide a comprehensive knowledge of each model's benefits and drawbacks. Finally, Grad-CAM visualizations are used to highlight the regions in CT images that influence the best model's predictions.

4.1 Performance Metrices

Our suggested deep learning models for multiclass stroke classification are evaluated using a variety of common performance criteria. **Accuracy**, which assesses the overall correctness of predictions. **Precision**, which is the ratio of true positives to all predicted positives. **Recall**, which calculates the percentage of true positives correctly identified out of all actual positives, are some examples of these metrics. The **F1-score**, which strikes a balance between recall and precision, is computed using both weighted (support-weighted) and macro (unweighted) averages. An especially important metric in medical diagnostics, where missed stroke cases can have major clinical repercussions, is the **False Negative Rate (FNR)**, which we also take into consideration to find the percentage of genuine positive cases that are mistakenly projected as negative. For each of the three categories—no stroke, hemorrhagic stroke, and ischemic stroke—we compute class-specific precision, recall, F1-score, and FNR in order to conduct a thorough assessment. In order to maintain class distribution across the folds and provide a reliable and objective evaluation of performance, we also use a **stratified 10-fold cross-validation** approach.

4.2 Results Analysis

The results show distinct variations in model performance (Tables 1 to 3). In Table 1, NS stands for No Stroke, HS for Hemorrhagic Stroke, and IS for Ischemic Stroke.

Although the bespoke CNN offer a helpful starting point, its comparatively shallow architecture resulted in low F1-scores and significant false negative rates (20–24%), suggesting a limited capacity to detect subtle hemorrhagic and ischemic characteristics.

While VGG19 with three layers unfrozen perform worse than with one layer, indicating that excessive fine-tuning can limit generalization, transfer learning

models consistently outperform the CNN: VGG16, VGG19, ResNet50, and ViT-Base improve with partial fine-tuning.

ResNet50, which benefited from residual connections and efficient adaption of pretrained features, perform the best, with F1-scores above 95% and false negatives below 5%. This decrease is clinically significant since false negatives are more harmful than false positives when diagnosing stroke.

ViT-Base also demonstrates competitive performance using its transformer architecture to capture long-range relationships over 16×16 picture patches through 12 encoder blocks, though, its performance is slightly lower than ResNet50 model, most likely as a result of the moderate dataset size since transformers usually need large training data.

Transfer learning models perform better overall than custom CNN because pretrained layers offer rich feature representations, deeper architectures capture intricate patterns, and layer-wise fine-tuning enables adaptation to stroke-specific CT properties. These patterns systematically show that for stroke classification, transfer learning is better than training a CNN from scratch.

The robustness of ResNet50 was further validated by cross-validation analysis (Table 4), which showed smooth training-validation curves (Figures 2) and minimal variation across folds, indicating consistent convergence and no overfitting. Strong separability is seen for all classes in the ROC–AUC curves (Figure 3), 99% for no stroke, 91% for bleeding, and 97% for ischemia.

The confusion matrix of the top two performers, ViT-Base and ResNet50 (Figure 4), on the test set illustrates the per-class predictions for no stroke, ischemic stroke, and hemorrhagic stroke. Finally, Figure 5 shows overall comparison bar charts for f1-score, recall, accuracy, and precision among models and class-wise comparison bar charts for FNR across models are shown in Figure 6.

In summary, ResNet50 with partial fine-tuning continuously performed better than all other models, providing stable results across folds, low FNR, and excellent accuracy. These results demonstrate how well our approach works to classify strokes from CT images in a way that is both dependable and clinically meaningful.

Finally, Table 5 compares the performance of our proposed model with that of existing methods documented in earlier research.

Table 1:
Class-wise Performance Metrics

Model	Class	Precision	Recall	F1-Score	FNR
CNN	**NS**	**0.947**	**0.934**	**0.940**	**0.066**
CNN	**HS**	**0.703**	**0.793**	**0.745**	**0.207**
CNN	**IS**	**0.822**	**0.763**	**0.791**	**0.237**
VGG16	**NS**	**0.952**	**0.950**	**0.951**	**0.050**
VGG16	**HS**	**0.800**	**0.835**	**0.816**	**0.165**
VGG16	**IS**	**0.858**	**0.823**	**0.840**	**0.177**
VGG16 (1 layer)	**NS**	**0.970**	**0.956**	**0.963**	**0.044**
VGG16 (1 layer)	**HS**	**0.800**	**0.854**	**0.830**	**0.146**
VGG16 (1 layer)	**IS**	**0.874**	**0.864**	**0.870**	**0.136**
VGG16 (2 layers)	**NS**	**0.974**	**0.964**	**0.969**	**0.036**
VGG16 (2 layers)	**HS**	**0.853**	**0.884**	**0.868**	**0.116**
VGG16 (2 layers)	**IS**	**0.894**	**0.899**	**0.897**	**0.101**
VGG16 (3 layers)	**NS**	**0.982**	**0.971**	**0.976**	**0.029**
VGG16 (3 layers)	**HS**	**0.904**	**0.915**	**0.909**	**0.085**
VGG16 (3 layers)	**IS**	**0.908**	**0.935**	**0.921**	**0.065**
VGG19 (1 layer)	**NS**	**0.974**	**0.977**	**0.976**	**0.023**
VGG19 (1 layer)	**HS**	**0.925**	**0.902**	**0.914**	**0.098**
VGG19 (1 layer)	**IS**	**0.906**	**0.917**	**0.912**	**0.083**
VGG19 (3 layers)	**NS**	**0.969**	**0.949**	**0.959**	**0.051**
VGG19 (3 layers)	**HS**	**0.833**	**0.884**	**0.858**	**0.116**
VGG19 (3 layers)	**IS**	**0.856**	**0.876**	**0.866**	**0.124**
ResNet50 (1 block)	**NS**	**0.989**	**0.982**	**0.986**	**0.018**
ResNet50 (1 block)	**HS**	**0.935**	**0.957**	**0.946**	**0.043**
ResNet50 (1 block)	**IS**	**0.947**	**0.953**	**0.950**	**0.047**
ViT-Base (1 block)	**NS**	**0.990**	**0.979**	**0.980**	**0.021**
ViT-Base (1 block)	**HS**	**0.920**	**0.930**	**0.920**	**0.070**
ViT-Base (1 block)	**IS**	**0.900**	**0.940**	**0.920**	**0.060**

Table 2:
Summary of Accuracy and Macro-Average Metrics

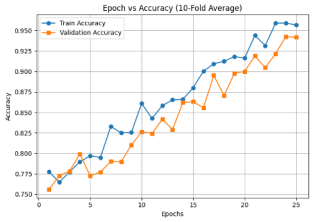
Model	Accuracy	Precision	Recall	F1-Score
CNN	**0.882**	**0.824**	**0.830**	**0.826**
VGG16	**0.910**	**0.869**	**0.869**	**0.869**
VGG16 (1 layer)	**0.924**	**0.881**	**0.891**	**0.886**
VGG16 (2 layer)	**0.940**	**0.907**	**0.916**	**0.911**
VGG16 (3 layer)	**0.956**	**0.931**	**0.94**	**0.936**
VGG19 (1 layer)	**0.955**	**0.935**	**0.932**	**0.934**
VGG19 (3 layer)	**0.926**	**0.886**	**0.903**	**0.894**
ResNet50 (1 block)	**0.973**	**0.957**	**0.964**	**0.961**
ViT-Base (1 block)	**0.960**	**0.930**	**0.950**	**0.940**

Table 3:
Weighted-Average Metrics

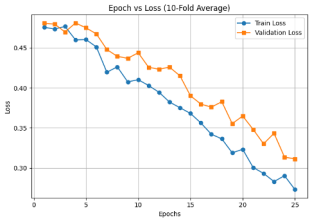
Model	Precision	Recall	F1-Score
CNN	0.881	0.882	0.880
VGG16	0.909	0.910	0.910
VGG16 (1 layer)	0.923	0.924	0.924
VGG16 (2 layer)	0.939	0.940	0.940
VGG16 (3 layer)	0.956	0.956	0.956
VGG19 (1 layer)	0.955	0.955	0.955
VGG19 (3 layer)	0.925	0.926	0.925
ResNet50 (1 block)	0.974	0.973	0.973
ViT-Base (1 block)	0.960	0.960	0.960

Table 4:
Fold-wise Metrics of ResNet50

Fold	Test Accuracy	Macro F1	Weighted F1
1	0.965	0.956	0.965
2	0.970	0.960	0.971
3	0.974	0.964	0.973
4	0.968	0.959	0.969
5	0.976	0.967	0.975
6	0.973	0.963	0.973
7	0.967	0.957	0.968
8	0.975	0.966	0.975
9	0.969	0.958	0.970
10	0.972	0.961	0.972
	Mean ± Std 0.973 ± 0.0034	Mean ± Std 0.961 ± 0.0032	Mean ± Std 0.973 ± 0.0033



(a) Epoch vs Accuracy (10-Fold Average) of ResNet50



(b) Epoch vs Loss (10-Fold Average) of ResNet50

Fig. 2: Training performance of ResNet50 across 10-fold cross-validation

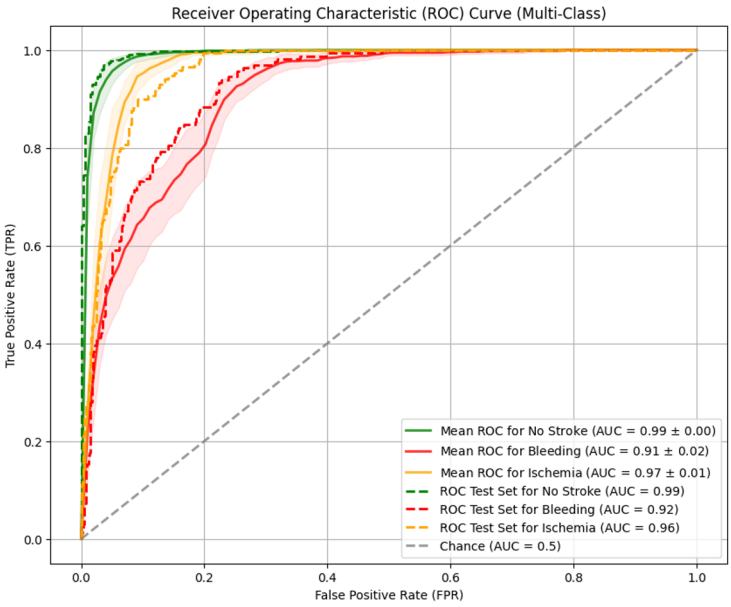
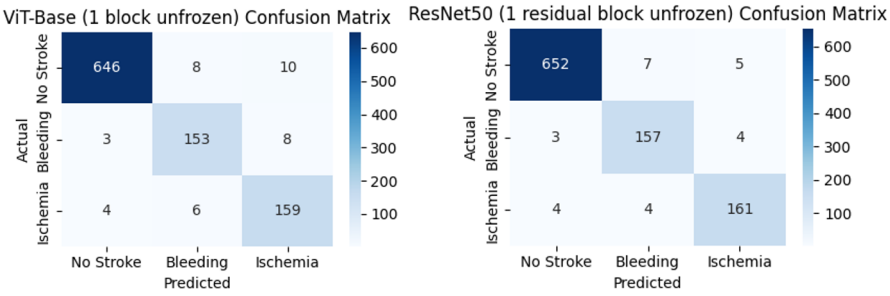


Fig. 3: ROC-AUC Curves of ResNet50



(a) ViT-Base (1 block unfrozen) Confusion Matrix

(b) ResNet50 (1 residual block unfrozen) Confusion Matrix

Fig. 4: Confusion matrices for ViT-Base and ResNet50 models

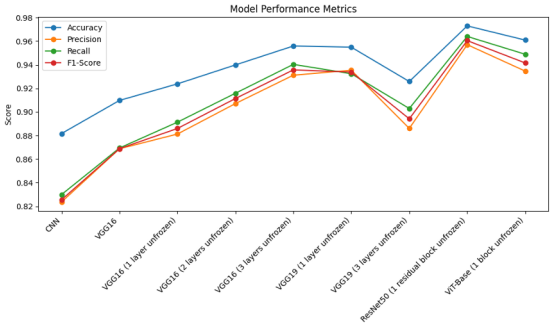


Fig. 5: Model Performance Metrics

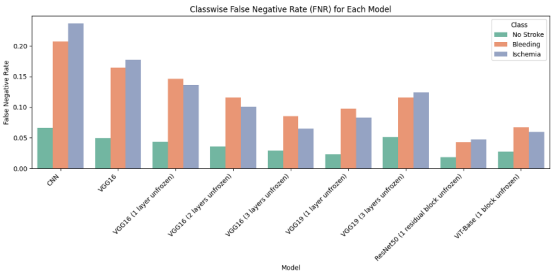


Fig. 6: Classwise FNR for each model

Table 5:
Performance Comparison With Existing Works

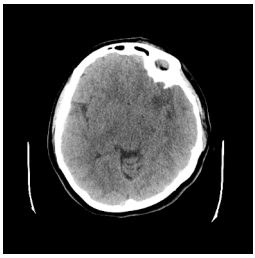
Paper	Accuracy	Precision	Recall	F1-Score	ROC-AUC
[13]	0.970	-	0.940	-	-
[15]	0.913	-	0.870	-	-
[16]	0.950	-	-	-	-
[19]	-	0.800	0.730	-	0.846
Our Paper	0.973	0.957	0.964	0.961	0.957 (Macro)

4.3 Grad-CAM Analysis

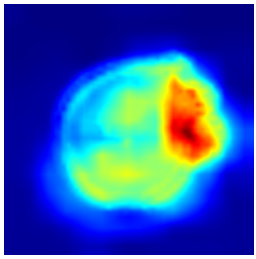
To evaluate the interpretability of the best model, Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to the ResNet50 outputs. Grad-CAM produces heatmaps that highlight the specific regions in the CT scans that contribute most strongly to the model's decision.

Figures 7 shows CT images and corresponding Grad-CAM visualizations for the different stroke categories. These examples indicate that the ResNet50 model consistently emphasizes anatomically and clinically relevant areas, confirming that its predictions are guided by meaningful radiological patterns. In the heatmaps, red and yellow tones represent regions with greater influence on the predicted class, whereas green and blue areas correspond to minimal or no contribution. This gradient of colors effectively illustrates the level of attention assigned to different image features.

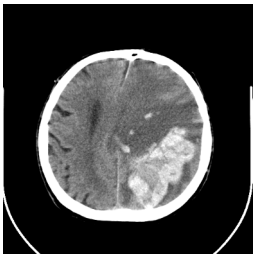
Overall, this interpretability assessment complements the quantitative evaluation and helps verify that the classification results are driven by coherent and clinically significant image characteristics, thereby improving clarity and confidence in the model outputs.



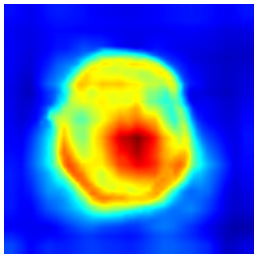
(a) Ischemia CT image



(b) Heatmap of ischemia CT image



(c) Bleeding CT image



(d) Heatmap of bleeding CT image

Fig. 7: CT images and corresponding Grad-CAM heatmaps

5 Conclusion

With the aim of proposing StrokeNetBench, a deep learning framework for multiclass stroke classification from CT images, this study evaluated a custom CNN alongside transfer learning models such as VGG16, VGG19, ResNet50, and ViT-Base. The incorporation of class weighting, stratified 10-fold cross-validation, and focal loss contributed to a strong and balanced performance, with ResNet50 achieving the highest accuracy and the lowest false-negative rates—an outcome of considerable clinical relevance.

Despite these strengths, several limitations remain. The analysis is based on a single dataset and uses only 2D CT slices rather than full volumetric scans. No patient demographic information was available, making it difficult to assess how well the models generalize across age groups, clinical histories, or other population-level variations. Additionally, the models were not validated on independent or multi-centre datasets, leaving uncertainty regarding their robustness across different scanners, acquisition protocols, and patient populations. The absence of multimodal imaging or accompanying clinical metadata further restricts comprehensive evaluation.

Future work will address these gaps by expanding validation to multi-centre and external datasets, integrating demographic and clinical variables to assess generalizability, and incorporating multimodal inputs to strengthen clinical applicability. Further development will also include the design of an intuitive user interface to support practical deployment in clinical settings. Overall, StrokeNet-Bench provides a strong foundation for advancing reliable and scalable tools for stroke assessment.

References

1. World Health Organization: The top 10 causes of death. (2024). Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
2. Feigin, V.L., Norrving, B., Mensah, G.A.: Global burden of stroke. *Circulation Research*, **120**(3), 439–448 (2017). doi:10.1161/CIRCRESAHA.116.308413
3. Aguirre, A.O., Rogers, J.L., Reardon, T., Shlobin, N.A., Ballatori, A.M., Brown, N.J., Gendreau, J., Shahrestani, S.: Stroke management and outcomes in low-income and lower-middle-income countries: a meta-analysis of 8535 patients. *Journal of Neurosurgery*, **139**(4), 1042–1051 (2023). doi:10.3171/2023.2.JNS222807
4. Powers, W.J., Rabinstein, A.A., Ackerson, T., Adeoye, O.M., Bambakidis, N.C., Becker, K., Biller, J., Brown, M., Demaerschalk, B.M., Hoh, B., et al.: Guidelines for the early management of patients with acute ischemic stroke: 2019 update. *Stroke*, **50**(12), e344–e418 (2019). doi:10.1161/STR.0000000000000211
5. Heit, J.J., Iv, M., Wintermark, M.: Imaging of intracranial hemorrhage. *Journal of Stroke*, **19**(1), 11 (2016). doi: 10.5853/jos.2016.00563
6. Saver, J.L.: Time is brain—quantified. *Stroke*, **37**(1), 263–266 (2006). doi:10.1161/01.STR.0000196957.55928.ab
7. Johnson, J.M., Lev, M.H.: Computed tomography in acute stroke. In: *The Stroke Book, Second Edition*, pp. 93–123. Cambridge University Press (2013). doi:10.1017/CBO9781139344296.009
8. Saba, L., Biswas, M., Kuppili, V., Cuadrado Godia, E., Suri, H.S., Edla, D.R., Omerzu, T., Laird, J.R., Khanna, N.N., Mavrogeni, S., et al.: The present and future of deep learning in radiology. *European Journal of Radiology*, **114**, 14–24 (2019). doi:10.1016/j.ejrad.2019.02.038
9. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Medical Image Analysis*, **42**, 60–88 (2017). doi:10.1016/j.media.2017.07.005
10. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, **542**(7639), 115–118 (2017)

11. Neethi, A.S., Niyas, S., Kannath, S.K., Mathew, J., Anzar, A.M., Rajan, J.: Stroke classification from computed tomography scans using 3D convolutional neural network. *Biomedical Signal Processing and Control*, **76**, 103720 (2022). doi:10.1016/j.bspc.2022.103720
12. Wang, Z., Yang, W., Li, Z., Rong, Z., Wang, X., Han, J., Ma, L.: A 25-year retrospective of the use of AI for diagnosing acute stroke: systematic review. *Journal of Medical Internet Research*, **26**, e59711 (2024). doi:10.2196/59711
13. UmaMaheswaran, S.K., Ahmad, F., Hegde, R., Alwakeel, A.M., Zahra, S.R.: Enhanced non-contrast computed tomography images for early acute stroke detection using machine learning approach. *Expert Systems with Applications*, **240**, 122559 (2024). doi:10.1016/j.eswa.2023.122559
14. Chauhan, S., Vig, L., De Filippo De Grazia, M., Corbetta, M., Ahmad, S., Zorzi, M.: A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images. *Frontiers in Neuroinformatics*, **13**, 53 (2019). doi:10.3389/fninf.2019.00053
15. Sahoo, P.K., Mohapatra, S., Wu, C.Y., Huang, K.L., Chang, T.Y., Lee, T.H.: Automatic identification of early ischemic lesions on non-contrast CT with deep learning approach. *Scientific Reports*, **12**(1), 18054 (2022)
16. Merdas, H.M.: Elastic Net–MLP–SMOTE (EMS)-Based model for enhancing stroke prediction. *Medinformatics*, **1**(2), 73–78 (2024). doi:10.47852/bonviewMEDIN42022470
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017)
18. Mundargi, Z., Khedkar, S., Kumbhar, S., Mohod, K., Meshram, Y.: Revolutionizing cerebral stroke prediction: mastery unveiled through stratified K-fold and K-fold cross validation techniques for imbalanced datasets. *Grenze International Journal of Engineering & Technology (GIJET)*, **10** (2024)
19. Arbabshirani, M.R., Fornwalt, B.K., Mongelluzzo, G.J., Suever, J.D., Geise, B.D., Patel, A.A., Moore, G.J.: Advanced machine learning in action: identification of intracranial hemorrhage on CT scans with clinical workflow integration. *NPJ Digital Medicine*, **1**(1), 9 (2018). doi:10.1038/s41746-017-0015-z
20. Eshmawi, A., Khayyat, M., Algarni, A.D., Hilali-Jaghdam, I.: An ensemble of deep learning enabled brain stroke classification model in magnetic resonance images. *Journal of Healthcare Engineering*, **2022**(1), 7815434 (2022). doi:10.1155/2022/7815434
21. Kaggle: Stroke CT Images Dataset. Available at: https://www.kaggle.com/datasets/ozguraslank/brain-stroke-ct-dataset. Accessed: 30 June 2025
22. Koç, U., Akçapınar Sezer, E., Özkaya, Y.A., et al.: Artificial Intelligence in Healthcare Competition (TEKNOFEST-2021): Stroke Data Set. *Eurasian Journal of Medicine*, **54**(3), 248–258 (2022). doi:10.5152/eurasianjmed.2022.22096