Application of Vision Transformer for Brain Stroke Classification Based on CT Images

Azhar Tursynova
Al-Farabi Kazakh National University, Kazakhstan | International Information Technology University, Kazakhstan
azhar.tursynova1@gmail.com (corresponding author)

Received: 30 June 2025 | Revised: 18 July 2025 and 27 August 2025 | Accepted: 2 September 2025

 DOI: https://doi.org/10.48084/etasr.13070

ABSTRACT

The development of artificial intelligence and machine learning has had a significant impact on medical diagnostics. This paper examines the application of the Vision Transformer (ViT) architecture for the task of classifying Computed Tomography (CT) images of the brain for the presence or absence of stroke signs. ViT which was originally developed for computer vision tasks, uses attention mechanisms, allowing the model to focus on important aspects of an image without first learning using task-specific data. The article presents the results of experiments on training the ViT model on a dataset of CT images, as well as performance comparisons with traditional methods such as Convolutional Neural Networks (CNNs). The results show that ViT can effectively classify strokes, demonstrating high accuracy and generalization ability based on new data. These findings may contribute to the wider application of transformer models in medical imaging which in the future may improve diagnostic accuracy and accelerate the treatment of stroke patients.

Keywords-brain stroke; vision transformer; CNN; CT; classification; DL

I. INTRODUCTION

Acute ischemic stroke delayed diagnosis and treatment can lead to serious consequences for brain function, while the risk of death increases. The validity of medical interventions, such as endovascular treatment, is determined by the location of the lesion in the posterior or anterior circulatory system and the period from the moment the disorder occurs. Most procedures in the treatment of patients with symptoms require the presence of human specialists. Contacting medical experts is a time-consuming process, and these specialists may not always be available in every medical facility. Imaging studies such as CT and Magnetic Resonance Imaging (MRI) are necessary to quickly diagnose brain damage in stroke and determine which areas of the parenchyma are damaged. Automated approaches to stroke assessment are needed to expand early treatment options [1].

Traditional methods of automatic identification and classification of cerebral infarcts have been developed using a set of recommendations for the design of functions provided by algorithm developers after a thorough analysis of clinical data. Since some aspects of a potential cerebral stroke are hidden and difficult to recognize on scans, traditional methods of automatic stroke classification have been hampered by insufficient complexity. On the other hand, deep learning methods allow the extraction of visual attributes from training samples, unlike conventional machine learning. These methods can simplify the modeling of cerebral infarcts and eliminate the limitations of previous approaches to deep learning. Convex kernels are used by Convolutional Neural Networks (CNNs) to extract certain features from an input image and to solve various image categorization tasks.

Modern artificial intelligence applications are designed to help solving a wide variety of problems. CNNs form one of the subcategories of deep learning that is currently widely used in neuroimaging [2]. Deep learning methods and the use of CNN are evaluated as a strategy for the diagnosis of acute ischemic strokes. Recently, the use of transformers has become one of the most popular research topics among the scientific community, which has produced amazing results in robotics, image recognition, and text recognition. Neural networks thrive in processing unstructured data, especially images, text, audio, and speech. Vision Transformers (ViTs) have proven to be best suited for handling such unstructured data [3].

II. RELATED WORK

Numerous studies have applied deep learning methods to stroke detection and classification, with CNN architectures being the most common approach. For example, authors in [4] implemented a CNN-based framework for early ischemic stroke detection, reporting an accuracy of 83% on CT data. More recent works have explored transfer learning techniques and hybrid models, combining CNNs with handcrafted features to improve performance. However, these approaches often require extensive preprocessing and struggle with generalization across datasets from different institutions. The Vision Transformer (ViT), initially introduced for natural image classification [5], has gained attention in medical imaging due to its ability to capture long-range dependencies in

data without convolutional layers. Applications of ViT in MRI-based brain lesion segmentation and classification have shown promising results, but its use in CT-based stroke detection remains limited. This study addresses this gap by evaluating the performance of ViT on CT brain scans for binary stroke classification.

III. METHODS

Simpler and more accessible diagnostic methods are Computed Tomography (CT) or MRI [1] with perfusion imaging. Ischemic stroke patient CT and MRI images can be seen in Figure 1.

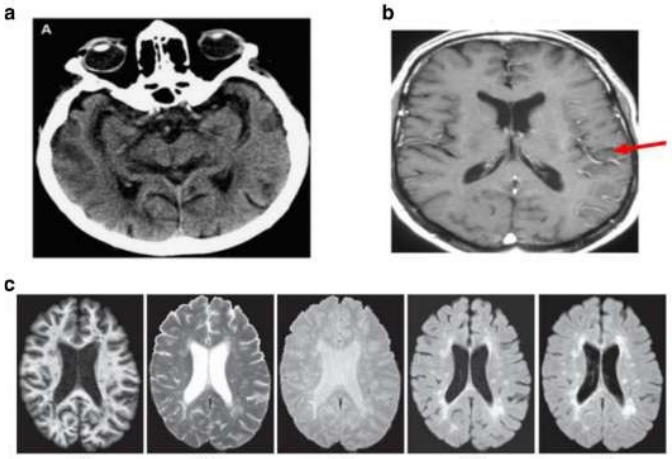

Fig. 1. (a) Non-contrast CT image showing transient ischemic attacks in the left half of the patient's body. (b) MRI image showing transient ischemic attacks in the right half of the patient's body. (c) Sample input image for the model using synthetic and fluid-attenuated inversion recovery (FLAIR) images.

A. Convolutional Neural Networks

Deep learning methods and the use of CNN are widely utilized in the diagnosis of acute ischemic strokes [4]. A popular topic in the field of automated diagnostics is the end-to-end architecture of the system. Recently, implemented algorithms for classifying brain strokes have been presented in several important publications. We implemented a compact end-to-end CNN baseline for binary stroke classification (stroke vs normal) on the CT images. The network comprises three convolutional blocks with 3×3 kernels, batch normalization, and ReLU activation function, each followed by 2×2 max-pooling. To mitigate overfitting, we apply dropout after the last block and before the classifier. Feature maps were aggregated by global average pooling and fed to a fully connected layer that outputs a single logit. Training uses binary cross-entropy with a sigmoid at inference. Input CT slices were rescaled to a fixed size and were intensity-normalized. Simple augmentations (random horizontal flip and small rotations within a clinically reasonable range) were used only on the training split. Adam optimizer was utilized (with initial learning rate equal to 9×10^{-5}) with a mini-batch size chosen to fit the GPU memory. The model was trained for up to 35 epochs with early stopping on validation loss. In our run, early stopping was triggered at epoch 14, and the best checkpoint (minimum validation loss) was used for all test-set results. Evaluation follows the dataset partitioning described above.

B. ViT for Image Classification

ViT for image classification is a model based on the ViT concept specifically designed for image classification tasks. This model is part of the Hugging Face Transformers collection and is a modified version of ViT, which is adapted for visual content analysis. The ViT architecture itself was first introduced in [5]. The authors adapted the technology of transformers, which were traditionally used for text processing, to work with images, allowing this model to analyze visual data with a new degree of efficiency.

For the transformer baseline we fine-tuned ViT-Base/16 (google/vit-base-patch16-224-in21k), using 16×16 patches, input size 224×224, an embedding dimension of 768, 12 transformer encoder layers, 12 attention heads, and an MLP with a hidden size of 3072, with a learnable class token and positional embeddings. The pre-trained backbone (ImageNet-21k) was followed by a 2-way classification head (normal, stroke). Images were resized and normalized via the model's default processor. Light geometric augmentations were applied only to the training split. We fine-tuned the model with AdamW optimizer (learning rate 2×10^{-5}, weight decay 0.01), the batch size was 16, for 5 epochs, monitoring validation loss in each epoch. The best checkpoint according to the validation loss was reported. At inference, the positive-class probability (stroke) was thresholded at 0.5.

C. Detailed Architecture Description

The convolutional network follows a classical feed-forward topology consisting of three convolutional stages (Conv–BN–ReLU–MaxPool) with a kernel size of 3 × 3 and increasing channel depth (32, 64, 128). Each stage performs spatial down-sampling via 2 × 2 max-pooling, gradually extracting higher-level spatial–semantic features. The final feature map is flattened through global average pooling and passed to a dense layer that produces a single activation corresponding to the stroke vs normal decision. Dropout layers (p = 0.25 and 0.5) are introduced to prevent overfitting. The model contains approximately 2.1 million trainable parameters. This compact design allows efficient training on limited medical datasets while maintaining sufficient representational capacity.

The Vision Transformer (ViT-Base/16) replaces the convolutional feature extractor with a transformer encoder operating on non-overlapping 16 × 16 image patches. Each patch is linearly projected into a 768-dimensional embedding and augmented with learnable positional encodings and a classification token. The transformer encoder comprises 12 identical blocks, each containing multi-head self-attention (12 heads) and a feed-forward network (hidden dimension 3072) with GELU activation. Layer normalization and residual connections are applied throughout. The final hidden state of the class token is passed to a two-unit linear head that outputs class logits (normal, stroke). This configuration contains about 86 million parameters and benefits from the global-context modeling typical of transformers, enabling improved discrimination of subtle stroke patterns.

D. Evaluation Metrics

Accuracy, precision, recall, and F1 score are the most common used metrics to evaluate forecasting results [6]. Accuracy is an indicator that evaluates a model's prediction capability in all parameters. It is measured as the percentage of correct predictions made by the model. It is defined by:

$$Accuracy(a) = \frac{\sum_{i=1}^{N}\left[a(x_i) = y_i\right]}{N} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where TP represent the true positive results, TN the true negative results, FP the false positive results, and FN the false negative results.

Precision provides an accurate representation of the validity of our positive results:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is a useful metric that can be used to try to accurately describe how our optimistic predictions correspond to the real world. It is defined by:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1 score is the average harmonic value between accuracy and Recall:

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (4)$$

IV. EXPERIMENTAL RESULTS

The publicly available Brain Stroke CT Image Dataset [7] was considered. It contains labeled CT brain images of both healthy individuals and stroke patients. This dataset was divided into three (training, validation, and testing). The training dataset contained 993 images of healthy people and 610 stroke cases, the validation dataset contained 240 images of healthy people and 146 stroke cases, and the training dataset contained 313 images of healthy people and 189 stroke cases. Some samples of the dataset can be seen in Figure 2.

The CNN model was trained for up to 35 epochs, and the results were: loss: 0.5575, accuracy: 0.7024, total_loss: 0.5602, validation accuracy: 0.7228, lr: 9.0000e-05. Figure 3 illustrates the graphs of loss and accuracy of the CNN model during training until the early stop at epoch 14. It can be seen that as the number of epochs increases, the accuracy of the model gradually increases, and the loss function decreases, which indicates that the model is well-trained. However, with a further increase in epochs, the effect of overfitting is possible, so the early stopping technique was used. As a result, the CNN model achieved a test accuracy of 81%. Figure 4 shows the confusion matrix for binary classification of cerebral stroke [8. 9]. In total, 313 normal cases and 189 stroke cases were included in the evaluation. The model correctly identified 273 normal images as "normal," and 134 stroke images as "stroke". However, there were 55 FN, where stroke images were misclassified as normal, and 40 FP, where normal images were incorrectly predicted as stroke. Based on these results, the calculated evaluation metrics are summarized in Table I.

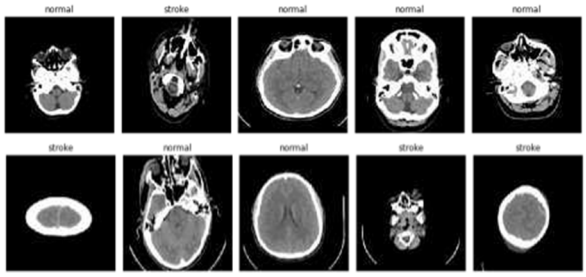

Fig. 2. Sample images from the Brain Stroke CT Image Dataset [8], illustrating healthy brain CT scans (left) and stroke-affected brain CT scans (right).

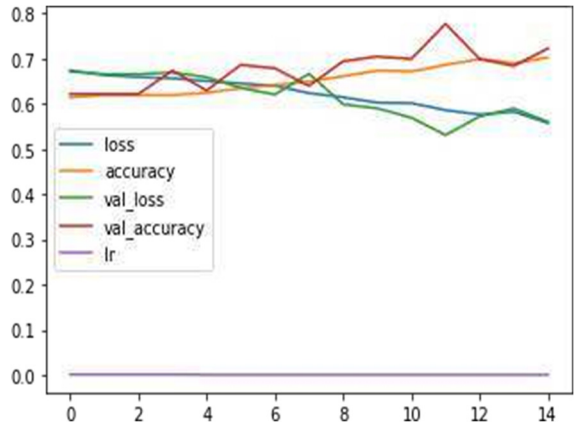

Fig. 3. The CNN model was trained for up to 35 epochs with early stopping triggered at epoch 14 to prevent overfitting.

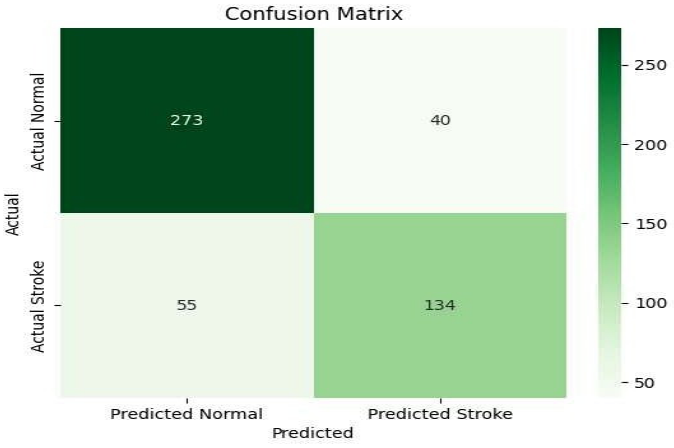

Fig. 4. Confusion matrix of the CNN model on the test set (N = 502).

The ViT model was trained for 5 epochs, and the results were: training loss = 0.007800, validation loss = 0.036258. The selection of the number of epochs for training the ViT model (5 epochs) was guided by the early stopping method. Training was stopped when the validation loss stabilized, indicating that the model had achieved an optimal level of accuracy. Table II shows the validation results.

TABLE I. RESULTS OF THE CT DATA TEST ON THE CNN MODEL

Metric	Value
Accuracy	0. 8108
Precision (Stroke)	0. 7701
Recall (Stroke)	0. 7090
F1-score (Stroke)	0. 7383
ROC-AUC	0. 875

The results of the confusion matrix for the ViT model are presented in Figure 5, which demonstrates noticeably fewer classification errors, especially FN cases, which are critical in medical diagnostics. The ViT achieved an accuracy of 95.2%, precision and recall of 0.94, and ROC-AUC of 0.99, indicating improved reliability in stroke detection compared with the CNN baseline. Both models were trained on the same training and validation splits. Due to computational constraints, the ViT model was evaluated on a stratified subset of the test data (N = 251), sampled with a fixed random seed to preserve the original class ratio, whereas the CNN model used the full test set (N = 502). This design does not affect the comparative conclusions, as distributional properties are matched and the evaluation protocol is otherwise identical. Figure 6 shows the ROC curves for the ViT model. The high area under both curves confirms the model's strong ability to distinguish stroke from normal cases, even under class imbalance conditions [10]. Figure 7 represents the Precision-Recall curve, which is especially important in the presence of unbalanced classes (in this case, the number of stroke images is less) [11]. The area under the curve confirm the model's stable ability to detect strokes even with high accuracy requirements.

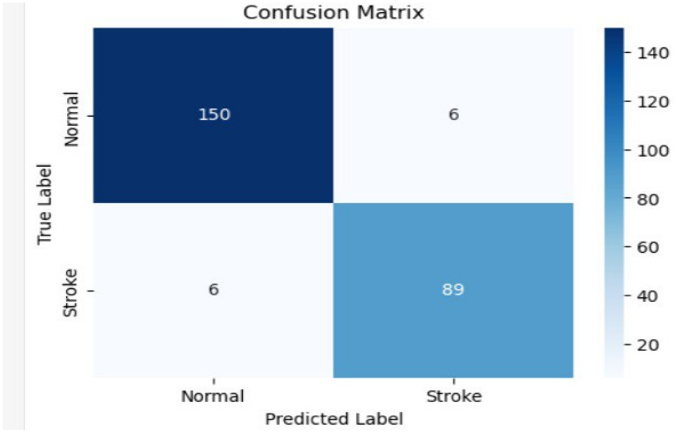

Fig. 5. Confusion matrix of the ViT model on a stratified test subset (N = 251).

TABLE II. RESULTS OF THE CT DATA TEST ON THE VIT MODEL

Metric	Value
Accuracy	0. 952191
Precision	0. 936842
Recall	0. 936842
F1-score	0. 936842
ROC-AUC	0. 994332

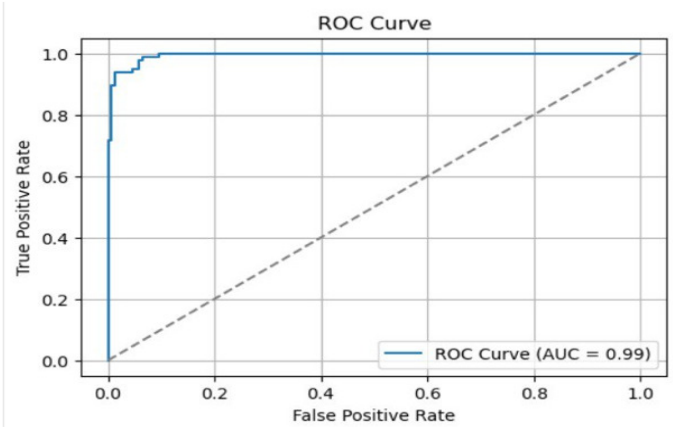

Fig. 6. ROC-AUC results of ViT model.

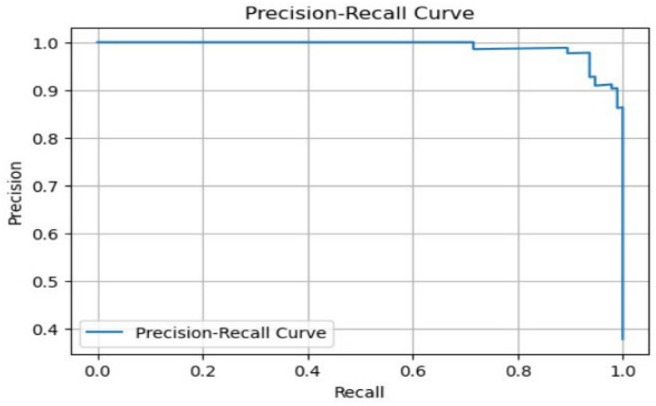

Fig. 7. Precision-Recall Curve results of ViT model.

The results of the ViT model surpass the results of the CNN model. Early stopping was used in both models. The ViT model was trained for 5 epochs, whereas the CNN model needed 14 epochs. Thus, based on the above results, it can be concluded that the use of transformers, such as ViT, opens up new prospects in the field of automated stroke diagnosis, increasing accuracy, reliability and speed of pathology detection from CT images. The training parameters, particularly the number of epochs, were determined based on preliminary experiments that monitored the validation loss metric. The early stopping technique was applied to prevent overfitting, resulting in optimal performance for the ViT model after significantly fewer epochs (5 epochs) compared to the CNN model (14 epochs).

V. DISCUSSION

The experimental results demonstrate that the ViT model significantly outperformed the conventional CNN approach in terms of accuracy, precision, recall, and F1-score, while also requiring fewer training epochs. These findings confirm the capability of transformer-based architectures to effectively capture both local and global contextual information in CT brain images, which is essential for accurate stroke detection.

From a practical perspective, the proposed ViT-based method has strong potential for integration into hospital Picture Archiving and Communication Systems (PACS) and

telemedicine platforms. Such integration can enable rapid pre-screening of CT scans for suspected stroke cases, reduce diagnostic delays, and minimize the occurrence of false negatives — a critical factor in time-sensitive medical conditions like the stroke. The method's efficiency and reduced computational requirements make it suitable for deployment even in medical facilities with limited computing resources or in remote areas where access to specialized neuroradiologists is scarce.

Despite these promising results, certain limitations should be acknowledged. The dataset used in this study, although publicly available and representative of stroke and non-stroke CT images, may not fully reflect the diversity of imaging protocols, scanner types, and patient demographics encountered in real-world clinical settings. Additionally, this work focuses on binary classification (stroke vs. no stroke) without differentiating between stroke subtypes such as ischemic and hemorrhagic, which may be clinically important for treatment planning. Another limitation is that the dataset size, while sufficient for initial validation, is still relatively small compared to large-scale medical imaging datasets used in other AI studies.

Future research should address these limitations by incorporating multi-center and multi-modal datasets, including MRI and perfusion imaging, to improve generalization. Furthermore, applying explainable AI techniques could enhance model interpretability, allowing clinicians to better understand the decision-making process of the ViT model and increasing trust in AI-assisted diagnostic systems.

VI. CONCLUSION

This study investigated the application of the Vision Transformer (ViT) architecture for brain stroke classification based on CT images and compared its performance with that of a traditional CNN model. The ViT model demonstrated superior results, achieving 95.2% accuracy and outperforming CNN in precision, recall, and F1-score, while requiring significantly fewer training epochs. As a resource-bounded setting, the ViT evaluation was performed on a stratified subset of the test set (N = 251). Future work will include full-set evaluation to confirm the robustness of these findings.

The novelty of this work lies in applying a transformer-based architecture to stroke classification using CT scans, which has been less explored in the literature compared to MRI-based approaches. This method offers improved diagnostic accuracy, reduced false negative rates, and faster model convergence, making it particularly valuable for clinical decision support in time-critical stroke diagnosis.

The contribution of this work includes: (1) adapting ViT to medical CT brain stroke classification tasks, (2) providing a performance comparison with CNN, and (3) demonstrating the feasibility of using transformer-based models for rapid and reliable stroke detection in real-world clinical settings.

Future work will focus on expanding the dataset to include multi-center and multi-modal imaging data, exploring explainable AI techniques for model interpretability, and integrating the approach into telemedicine systems to assist in remote stroke diagnosis.

ACKNOWLEDGMENT

This work was supported by the research project "A comprehensive system for diagnosing brain stroke using artificial intelligence funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No.IRN AP22686812." The supervisor of the project is Azhar Tursynova.

REFERENCES

[1] P. Vilela and H. A. Rowley, "Brain ischemia: CT and MRI techniques in acute ischemic stroke," *European Journal of Radiology*, vol. 96, pp. 162–172, Nov. 2017, https://doi.org/10.1016/j.ejrad.2017.08.014.

[2] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, Mar. 2021, Art. no. 53, https://doi.org/10.1186/s40537-021-00444-8.

[3] Y. Liu *et al.*, "A Survey of Visual Transformers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7478–7498, Jun. 2024, https://doi.org/10.1109/TNNLS.2022.3227717.

[4] C.-L. Chin *et al.*, "An automated early ischemic stroke detection system using CNN deep learning algorithm," in *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, Taichung, Taiwan, Aug. 2017, pp. 368–372, https://doi.org/10.1109/ICAwST.2017.8256481.

[5] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021, https://doi.org/10.48550/arXiv.2010.11929.

[6] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Online, Aug. 2020, pp. 79–91, https://doi.org/10.18653/v1/2020.eval4nlp-1.9.

[7] A. Rahman, "Brain Stroke CT Image Dataset." 2021, [Online]. Available: https://www.kaggle.com/datasets/afridirahman/brain-stroke-ct-image-dataset.

[8] A. Tursynova *et al.*, "Deep Learning-Enabled Brain Stroke Classification on Computed Tomography Images," *Computers, Materials and Continua*, vol. 75, no. 1, pp. 1431–1446, Jan. 2023, https://doi.org/10.32604/cmc.2023.034400.

[9] T. Rohini and P. Praveen, "An Intuitive Approach on Transfer Learning with an IBF+IHP Model for Stroke Classification and Prediction," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19655–19660, Feb. 2025, https://doi.org/10.48084/etasr.9031.

[10] M. H. Ferris *et al.*, "Using ROC curves and AUC to evaluate performance of no-reference image fusion metrics," in *2015 National Aerospace and Electronics Conference (NAECON)*, Dayton, OH, USA, Jun. 2015, pp. 27–34, https://doi.org/10.1109/NAECON.2015.7443034.

[11] J. Miao and W. Zhu, "Precision–recall curve (PRC) classification trees," *Evolutionary Intelligence*, vol. 15, no. 3, pp. 1545–1569, Sep. 2022, https://doi.org/10.1007/s12065-021-00565-2.