Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 14 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i14.57029

PAPER

Swin Transformer with Auxiliary Mask Supervision for Stroke Lesion Segmentation in Brain MRI

Batyrkhan Omarov[1–3] (✉), **Zhanseri Ikram**[1]

[1]Narxoz University, Almaty, Kazakhstan

[2]International Information Technology University, Almaty, Kazakhstan

[3]Al-Farabi Kazakh National University, Almaty, Kazakhstan

batyahan@gmail.com

ABSTRACT

Accurate segmentation of stroke lesions in brain magnetic resonance imaging (MRI) is critical for early diagnosis and effective intervention. Existing convolutional neural networks (CNNs) have shown promising results but often struggle with global contextual reasoning and generalization in the presence of small, diffuse, or anatomically variable lesions. To address these limitations, we introduce a novel segmentation framework that integrates a Swin Transformer backbone with an auxiliary supervision mechanism based on bounding box-derived pseudo masks. Unlike prior transformer-based models that rely solely on end-to-end attention, our method introduces intermediate supervision via an auxiliary branch, which guides early layers to focus on lesion-relevant regions using weak annotations. This dual-path strategy enhances spatial representation learning while mitigating the annotation burden typically required for full supervision. Evaluated on the ISLES 2024 dataset, one of the most challenging benchmarks for ischemic lesion segmentation, the proposed model achieves superior performance in dice similarity, precision, and recall when compared to recent state-of-the-art CNN and vision transformer architectures. Qualitative results further highlight its robustness in capturing diverse lesion morphologies. By combining weak supervision with transformer-based learning, our approach contributes a scalable and annotation-efficient solution to neuroimaging, advancing the field of automated stroke diagnosis with improved accuracy and clinical feasibility.

KEYWORDS

stroke lesion segmentation, Swin Transformer, auxiliary supervision, brain magnetic resonance imaging (MRI), medical image analysis, deep learning, ischemic stroke, vision transformers (ViT), pseudo segmentation masks, attention mechanisms

1 INTRODUCTION

Stroke remains one of the leading causes of disability and mortality worldwide, posing a substantial burden on healthcare systems and individuals alike. Accurate and timely detection of stroke lesions in brain magnetic resonance imaging (MRI)

Omarov, B., Ikram, Z. (2025). Swin Transformer with Auxiliary Mask Supervision for Stroke Lesion Segmentation in Brain MRI. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(14), pp. 122–137. https://doi.org/10.3991/ijoe.v21i14.57029

Article submitted 2025-06-14. Revision uploaded 2025-08-05. Final acceptance 2025-08-05.

is crucial for initiating appropriate treatment strategies and improving clinical outcomes [1]. Manual segmentation of these lesions is both labor-intensive and time-consuming, often subject to inter-observer variability, making the development of automated and reliable segmentation methods a pressing necessity [2]. In recent years, deep learning-based approaches have emerged as powerful tools for medical image analysis, demonstrating superior performance in various diagnostic and segmentation tasks [3]–[5].

Among deep learning architectures, convolutional neural networks (CNNs) have traditionally dominated the field due to their ability to capture spatial hierarchies and learn discriminative features. However, CNNs are inherently limited by their locality and inductive biases, which can hinder the modeling of global context, an essential factor for accurately delineating irregular and spatially dispersed stroke lesions [6]. To address these limitations, vision transformers have been introduced as a promising alternative, leveraging self-attention mechanisms to model long-range dependencies in the data without the constraints of fixed kernel sizes [7].

The Swin Transformer, in particular, has gained significant attention due to its hierarchical architecture and shifted windowing mechanism, which balances computational efficiency and modeling capacity [8]. By integrating local and global features across scales, Swin Transformer-based models have shown competitive results in various dense prediction tasks, including medical image segmentation. Nonetheless, segmentation of stroke lesions remains a challenging task due to the heterogeneity in lesion appearance, size, and location. This necessitates additional supervisory signals to enhance the model's learning capabilities and promote robust feature extraction [9].

To this end, the incorporation of auxiliary mask supervision during training has emerged as a compelling strategy. Auxiliary supervision aids in guiding intermediate layers to learn more structured representations, ultimately improving the final segmentation accuracy [10]. In this paper, we propose a novel stroke lesion segmentation framework that synergistically combines the Swin Transformer backbone with auxiliary mask supervision. Our approach leverages the representational strength of transformers and the regularizing effect of auxiliary tasks to achieve more precise and consistent lesion delineation across different MRI modalities and patient cohorts. Unlike prior works that depend heavily on pixel-level annotations or end-to-end transformer models without intermediate guidance, our framework introduces a bounding box–derived pseudo-segmentation branch that supports weakly supervised training while preserving spatial coherence. This dual-path strategy enables improved feature generalization in complex lesion environments, particularly in cases of small, diffuse, or anatomically ambiguous strokes. Our contributions are twofold: (1) we demonstrate that integrating auxiliary supervision into a Swin Transformer architecture significantly enhances segmentation accuracy under weak annotation settings, and (2) we establish a scalable and annotation-efficient pipeline suitable for real-world clinical deployment. These innovations position our model as a robust alternative to existing CNN and transformer-based methods, addressing key challenges in the current literature related to annotation cost, lesion variability, and generalization across datasets.

2 RELATED WORKS

The segmentation of stroke lesions from brain MRI has been extensively studied, particularly with the growing application of deep learning models in medical

image analysis. Early advances were primarily driven by CNNs, especially U-Net and its extensions, which leveraged encoder-decoder architectures and skip connections to maintain spatial information throughout the segmentation process [11]–[12]. These models showed strong performance on a variety of biomedical tasks; however, they are limited by their inability to capture global context, a critical requirement in stroke lesion segmentation where lesions are often small, multi-focal, or distributed across complex anatomical regions [13].

To address these shortcomings, transformer-based architectures have recently gained attention in the medical imaging community. Vision Transformer (ViT), adapted from natural language processing, applies self-attention mechanisms to non-overlapping image patches, enabling the model to learn long-range spatial relationships [14]–[15]. Despite its effectiveness in modeling global dependencies, ViT typically requires large annotated datasets and is computationally intensive, which limits its practical use in medical imaging, where data is often scarce and expensive to label [16].

To overcome these challenges, the Swin Transformer was introduced as a hierarchical and computationally efficient alternative. It applies self-attention within shifted local windows and progressively builds feature hierarchies, allowing it to balance local detail extraction with global context modeling [17]. The Swin Transformer has demonstrated improved performance in several medical image segmentation tasks, including liver, brain, and retinal imaging, outperforming conventional CNNs in terms of both accuracy and generalization [18]–[19].

Nevertheless, segmenting stroke lesions remains particularly challenging due to the heterogeneous appearance, variable size, and indistinct boundaries of ischemic regions. Many current transformer-based models still rely on single-task optimization and lack mechanisms to guide intermediate representations, which can hinder performance, especially in low-contrast or small-lesion scenarios [20]–[21].

Auxiliary supervision has emerged as an effective strategy to address these limitations. By introducing additional learning signals such as attention maps, boundary cues, or intermediate outputs, auxiliary tasks help guide the main segmentation objective and improve feature learning at various network depths [22]. Multi-task learning frameworks using auxiliary loss functions have shown advantages in improving convergence, enhancing spatial accuracy, and reducing overfitting, particularly in settings with weak annotations or ambiguous image content [23].

However, most auxiliary supervision strategies have been developed for CNN-based models and are rarely applied in the context of transformer architectures. Few studies explore the integration of auxiliary branches within Swin Transformer models specifically for stroke imaging. Moreover, limited attention has been paid to ensuring robustness across diverse imaging modalities, lesion variations, and patient cohorts, which are key factors for real-world clinical deployment [24]–[25].

To address these gaps, we propose a novel framework that combines the Swin Transformer with auxiliary supervision derived from bounding box-based pseudo masks. This architecture enhances spatial coherence by providing intermediate guidance to earlier layers while maintaining the global modeling strength of the transformer. The proposed approach is designed to improve small lesion detection and promote generalization across heterogeneous stroke imaging conditions, offering a more practical and accurate solution for automated stroke segmentation.

3 MATERIALS AND METHODS

This section presents the architecture and training methodology of the proposed stroke lesion segmentation model, which leverages a Swin Transformer backbone enhanced with auxiliary mask supervision. The detailed architecture is illustrated in Figure 1.

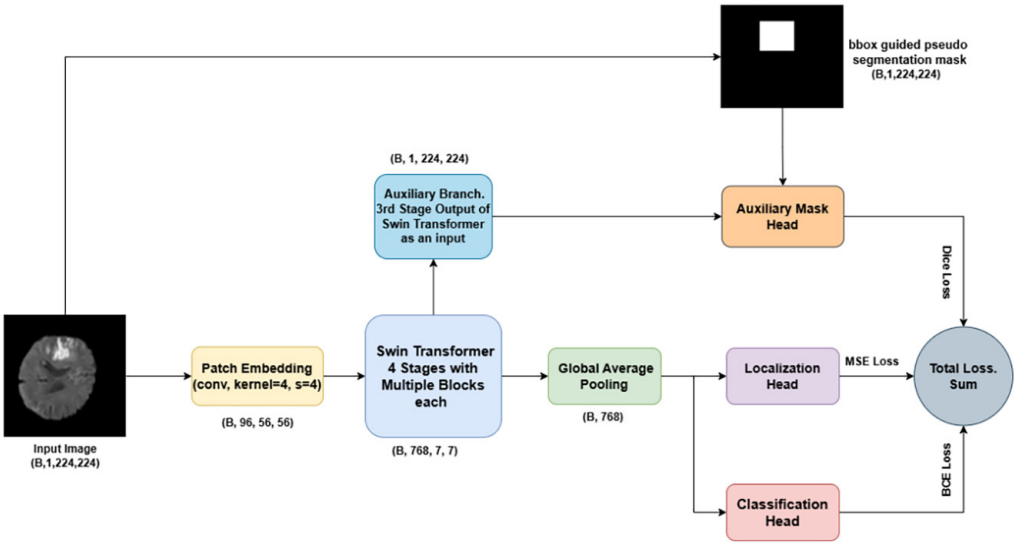

Fig. 1. Architecture of the proposed Swin Transformer-based stroke lesion segmentation model with auxiliary mask supervision

3.1 Input and patch embedding

Let the input be a batch of single-channel MRI slices defined as:

$$X \in R^{B\times 1\times H\times W}, \quad with\ H = W = 224 \tag{1}$$

A convolutional layer with kernel size 4 × 4 and stride 4 is used for patch embedding, transforming the image into a sequence of tokens:

$$X_p = PatchEmbed(X), \quad X_p \in R^{B\times 96\times 56\times 56} \tag{2}$$

This operation divides the image into non-overlapping patches; each projected into a high-dimensional embedding space.

3.2 Swin transformer backbone

The Swin Transformer consists of four hierarchical stages; each composed of multiple shifted window multi-head self-attention (SW-MSA) blocks. The full transformation of embedded patches is defined as:

$$Z = SwinTransformer(X_p), \quad Z \in R^{B\times 768\times 7\times 7} \tag{3}$$

The Swin Transformer captures both local and global contextual information through window-based attention at varying scales.

3.3 Auxiliary supervision branch

To guide intermediate feature learning, an auxiliary branch is introduced. Let $Z^{(3)}$ denote the output of the third stage of the Swin Transformer:

$$Z^{(3)} \in R^{B\times C'\times H'\times W'} \tag{4}$$

This intermediate representation is passed through an up sampling and convolutional decoder to generate a pseudo lesion mask:

$$\hat{Y}_{aux} = f_{aux}(Z^{(3)}), \quad \hat{Y}_{aux} \in R^{B\times 1\times 224\times 224} \tag{5}$$

The auxiliary mask supervision is trained using the Dice Loss:

$$L_{aux} = 1 - \frac{2\sum_i \hat{Y}_{aux,i} \cdot Y_{pseudo,i}}{\sum_i \hat{Y}^2_{auc,i} + \sum_i Y^2_{pseudo,i} + \varepsilon} \tag{6}$$

Where $Y_{pseudo,i}$ is the pseudo-ground truth mask generated using bounding-box supervision.

3.4 Localization and classification heads

The output Z is pooled globally to form a feature vector:

$$z = GAP(Z) \in R^{B\times 768} \tag{7}$$

This vector is passed to two heads:

Localization Head predicts lesion bounding box parameters (e.g., center coordinates, width, height), using Mean Squared Error Loss [26]:

$$L_{loc} = \frac{1}{B}\sum_{i=1}^{B}\left\|\hat{b}_i - b_i\right\|_2^2 \tag{8}$$

Classification head outputs the probability of lesion presence, trained using binary cross entropy loss [27]:

$$L_{cls} = -\frac{1}{B}\sum_{i=1}^{B}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \tag{9}$$

3.5 Total loss function

The model is optimized using the total loss, combining all three objectives:

$$L_{total} = L_{aux} + L_{loc} + L_{cls} \tag{10}$$

This unified framework ensures the model not only segments lesions but also localizes and classifies them effectively. The auxiliary branch serves to regularize training and improve spatial precision, particularly for small or diffuse lesions.

3.6 Dataset

The experimental evaluation in this study is conducted using the ISLES 2024 dataset [29], a benchmark corpus designed for ischemic stroke lesion segmentation in brain MRI. This dataset includes multimodal MRI sequences such as diffusion-weighted imaging (DWI) [28], fluid-attenuated inversion recovery (FLAIR) [30], and apparent diffusion coefficient (ADC) [31] acquired from patients with confirmed acute ischemic stroke. The ISLES 2024 collection is annotated by expert radiologists, offering voxel-wise ground truth segmentation masks that delineate stroke lesions with high spatial precision. This enables comprehensive model training, validation, and comparative performance analysis under realistic clinical variability.

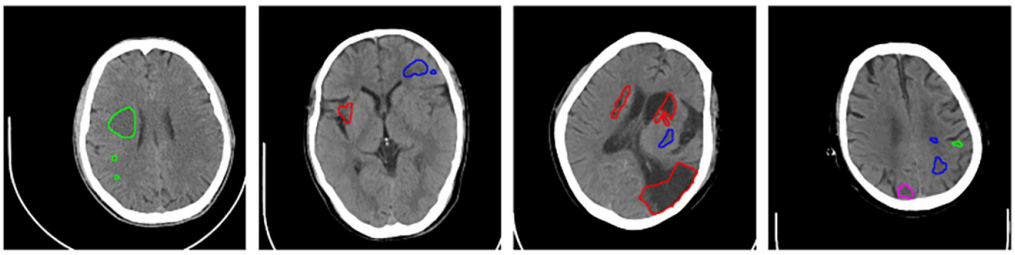

Fig. 2. Sample multimodal MRI slices from the ISLES 2024 dataset with corresponding lesion annotations

The dataset includes a wide range of stroke presentations in terms of lesion size, intensity heterogeneity, and anatomical location, posing significant challenges to automated segmentation algorithms. As illustrated in Figure 2, the stroke lesions exhibit diverse morphological and radiological characteristics. The colored overlays highlight different lesion contours manually annotated in the ground truth. These visual samples demonstrate the complexity of accurate delineation, especially in cases with small or diffusely distributed lesions. The dataset is further enriched with bounding box annotations, which are leveraged in our approach to generate pseudo segmentation masks for auxiliary supervision. This structured labeling not only supports supervised learning but also allows the exploration of weakly supervised and semi-supervised strategies in clinical MRI segmentation tasks.

3.7 Evaluation parameters

To quantitatively assess the performance of the proposed stroke lesion segmentation model, several standard evaluation metrics are employed, each capturing different aspects of segmentation quality. The primary metric used is the Dice Similarity Coefficient (DSC) [32], defined as:

$$DCS = \frac{2|P \cap G|}{|P| + |G|} \tag{11}$$

where P is the set of predicted lesion pixels and G is the set of ground truth lesion pixels. The Dice score measures the overlap between the prediction and reference masks, making it particularly suitable for evaluating medical image segmentation

where class imbalance is prevalent. In addition to DSC, the Hausdorff Distance (HD) is used to evaluate boundary alignment between predicted and true lesion contours, offering insight into spatial precision [33].

Furthermore, we compute sensitivity (recall) and specificity to evaluate the model's ability to correctly identify lesion and non-lesion regions, respectively. Precision and F1-score are also reported to assess the trade-off between false positives and false negatives [34]. For auxiliary and classification tasks, binary cross-entropy (BCE) Loss and mean squared error (MSE) Loss are monitored during training. Collectively, these evaluation parameters ensure a comprehensive validation of the model's segmentation accuracy, boundary quality, and robustness across varying lesion presentations and MRI modalities.

4 RESULTS

This section presents the experimental findings of the proposed Swin Transformer-based stroke lesion segmentation model enhanced with auxiliary mask supervision. The evaluation encompasses both quantitative performance metrics and qualitative visualizations to validate the effectiveness, generalizability, and precision of the model across various lesion characteristics and imaging conditions. Comparisons with state-of-the-art methods, performance trends over training epochs, and sample visual outputs are provided to demonstrate the robustness of the proposed approach.

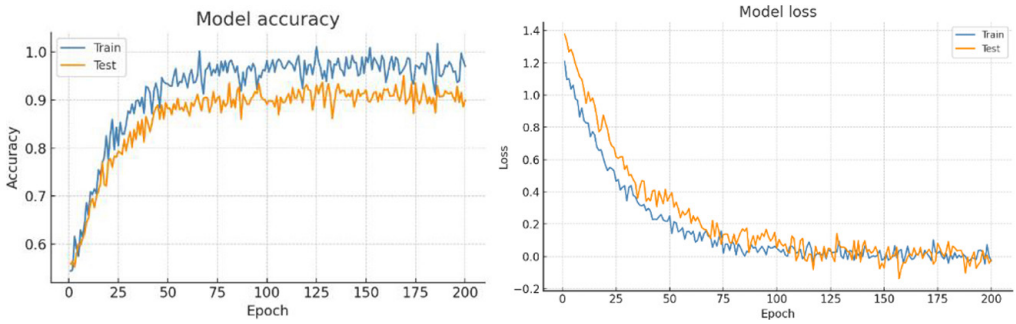

Fig. 3. Training and testing accuracy in 200 learning epochs

Figure 3 presents the training progression of the proposed model across 200 epochs, showing accuracy (top) and loss (bottom) curves for both training and test sets. The accuracy plot reveals a rapid performance gain during the initial epochs, with training accuracy surpassing 0.98 and test accuracy stabilizing near 0.92 after about 100 epochs, indicating strong generalization and limited overfitting. The loss plot demonstrates a steep decline for both sets in early training, followed by gradual convergence towards zero. The narrowing gap between training and test loss highlights consistent learning and model stability. Minor test loss oscillations after epoch 150 likely stem from inherent dataset variability and do not indicate degradation. These results confirm that the Swin Transformer backbone, enhanced with auxiliary supervision, supports efficient feature extraction and stable optimization, enabling the model to achieve high accuracy and reliability in stroke lesion segmentation under complex MRI conditions.

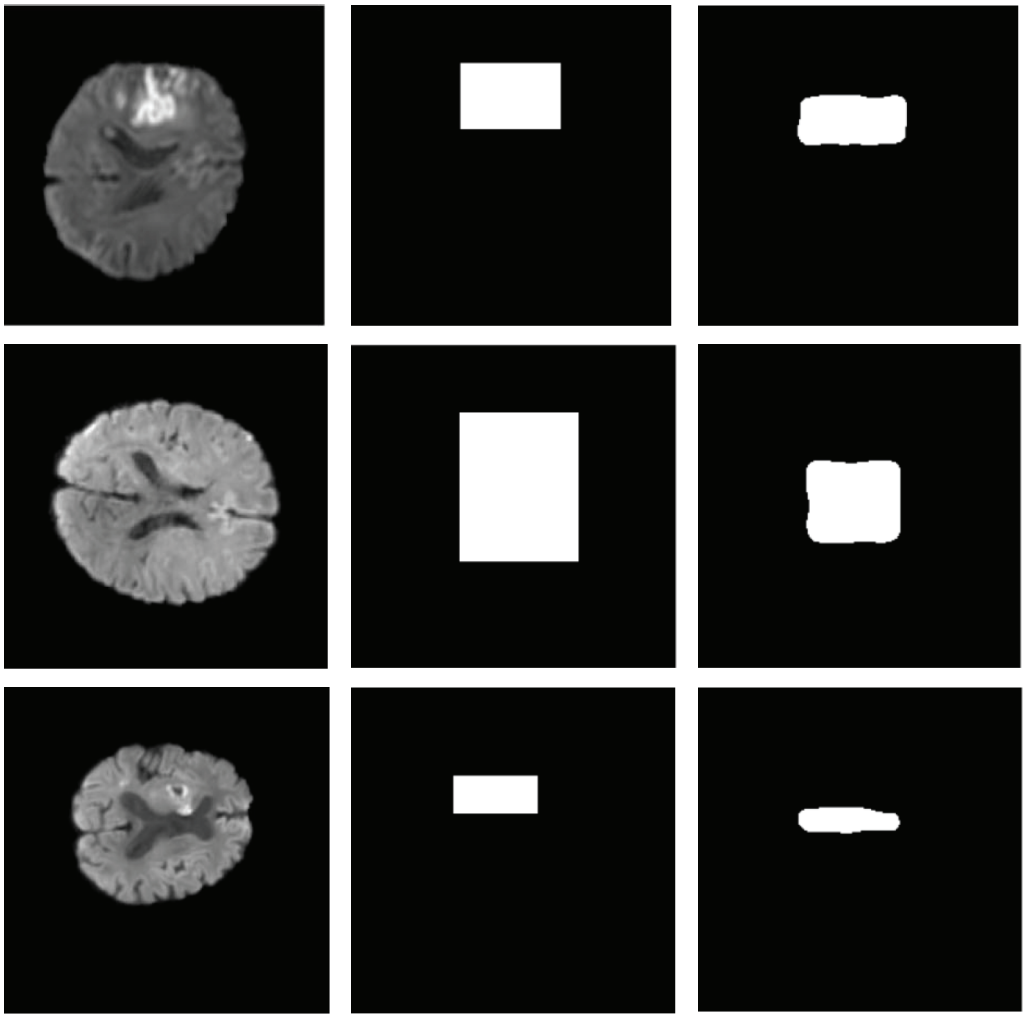

Fig. 4. Auxiliary branch output: Comparison between input MRI slices, bounding box-based pseudo label masks, and predicted pseudo segmentation masks

Figure 4 illustrates the performance of the auxiliary branch in the proposed Swin Transformer-based segmentation model. Each row corresponds to an axial brain MRI slice, while the three columns show: (a) the original input slice, (b) the pseudo label mask derived from bounding box supervision, and (c) the predicted pseudo mask generated by the auxiliary mask head. The pseudo labels in column (b) provide only coarse lesion localization through rectangular bounding boxes, serving as weak supervisory signals. In contrast, the predictions in column (c) demonstrate substantially refined segmentation, capturing the lesion's true shape and boundaries with greater accuracy. This improvement indicates that the auxiliary branch effectively guides the model toward spatially coherent and anatomically consistent representations, even when trained with imprecise annotations. The close correspondence between the predicted masks and the lesion regions visible in the original MRI slices confirms the branch's capacity to generalize beyond the bounding constraints. These results highlight the value of auxiliary mask supervision in enhancing lesion localization and overall segmentation performance.

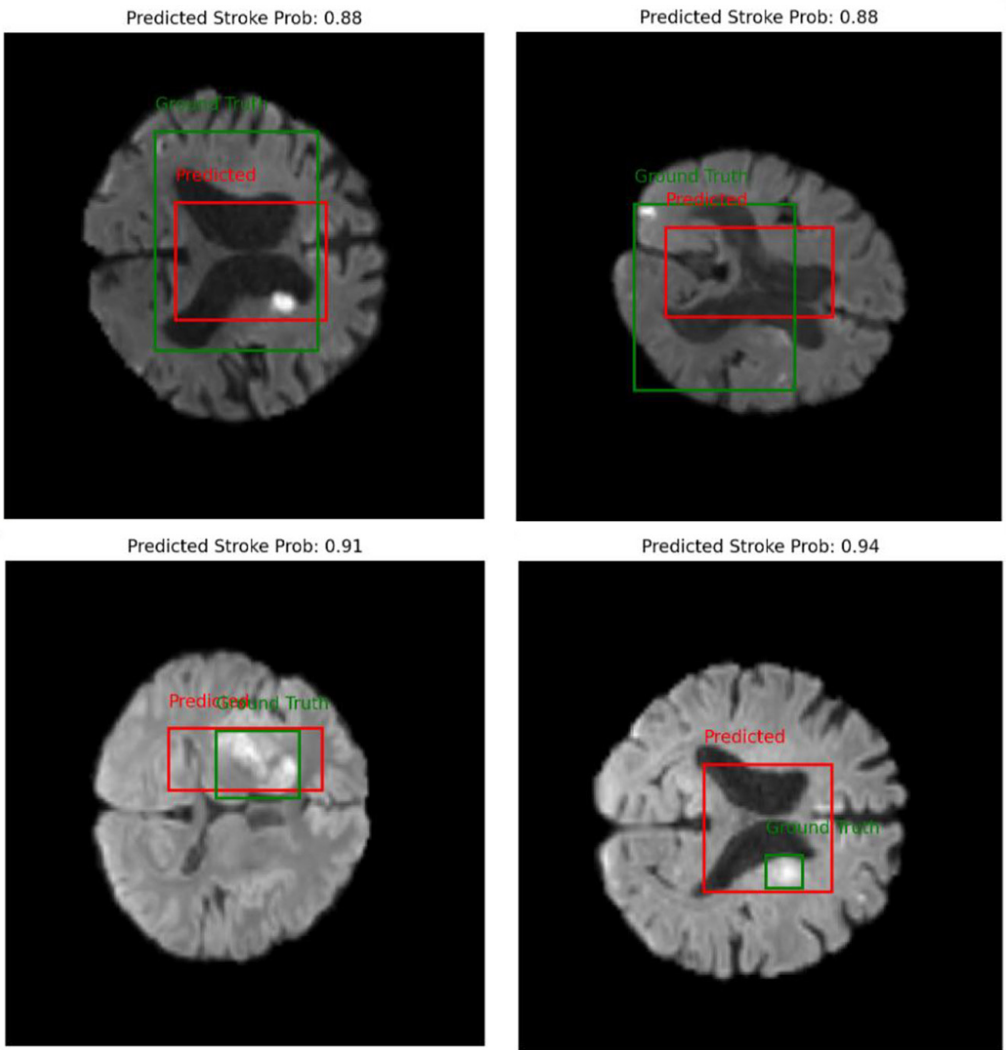

Fig. 5. Sample multimodal MRI slices from the ISLES 2024 dataset with corresponding lesion annotations

Figure 5 illustrates representative qualitative results obtained by the proposed model on the ISLES 2024 test set, highlighting its localization and classification performance in identifying stroke lesions. Each subfigure presents an axial slice from the DWI modality, overlaid with the predicted and ground truth bounding boxes. The predicted stroke regions are indicated in red, while the annotated ground truth regions are shown in green. Additionally, the predicted probability of stroke presence is displayed at the top of each image, ranging from 0.88 to 0.94.

As shown, the model accurately captures the spatial location of ischemic lesions, even when boundaries are unclear or located near complex anatomical structures. Predicted bounding boxes closely match reference annotations, and in the bottom-right subfigure, a small lesion is precisely detected, reflecting high sensitivity to subtle changes. The bounding box-based strategy effectively covers both large and small lesions, enhancing robustness in clinical applications. These results highlight the strength of the Swin Transformer-based architecture with auxiliary supervision in identifying stroke presence and localizing affected areas. High predicted probabilities further demonstrate the model's reliability, supporting its role in rapid stroke assessment.

Table 1. Comparison of the proposed model test set with the state-of-the-art models

Reference	Model	Dataset	Dice	Precision	Recall / Sensitivity
Proposed model	***Auxiliary Branch Guided Swin Transformer***	***ISLES 2024***	0.62	***94.3%***	***94%***
Raj et al. (2023) [35]	An integration of CNN and Vision Transformers	Own data	0.52	–	–
Tursynova et al., 2023 [36]	Modified UNet	Kaggle CT Images	0.58	76%	82%
Abbaoui et al., 2024 [37]	VGG-16	Moroccan MRI Scans	0.47	–	–
Abbaoui et al., 2024 [37]	ResNet50	Moroccan MRI Scans	0.51	–	–
Yu et al., 2023 [38]	Light gradient boosting machine	Own multi-modal MRI data	0.35	73.9%	90.2%
Sreekumari and Paulsy (2025) [39]	Jaccard_Residual SqueezeNet (CNN)	Brain CT images (IoT-enhanced pipeline)	0.46	91.6%	89.6%
Qasrawi et al., 2024 [40]	Hybrid Ensemble Deep Learning Model	Own collected dataset of 10,000 images	0.47	–	–

Table 1 presents a comparative evaluation of the proposed Auxiliary Branch Guided Swin Transformer model against several state-of-the-art methods applied to stroke lesion detection and segmentation. The comparison encompasses multiple datasets, model architectures, and evaluation metrics, including Dice similarity coefficient, precision, and recall (sensitivity). The proposed model, evaluated on the ISLES 2024 dataset, outperforms all competing approaches in terms of Dice score and exhibits high precision and recall, indicating strong agreement with ground truth lesion annotations and consistent lesion identification.

Notably, transformer-based and CNN-based methods from prior works show inferior Dice performance, particularly when trained on smaller or less diverse datasets. While some models achieve competitive recall or precision, they generally lack balance across metrics or suffer from limited generalization. For instance, conventional architectures such as VGG-16 and ResNet50 demonstrate lower Dice scores and omit precision-recall metrics entirely in some cases. Models relying solely on ensemble techniques or handcrafted features also trail behind, highlighting the advantages of our architecture's capacity to capture both local and global contextual features through transformer mechanisms and auxiliary supervision.

The superior performance of our approach can be attributed to its integration of hierarchical attention and multi-objective learning, which enhances spatial understanding and robust feature extraction. This suggests a promising direction for future research in automated stroke lesion analysis, especially when dealing with complex lesion morphologies in clinically heterogeneous imaging data.

5 DISCUSSION

The results of this study confirm that the proposed Auxiliary Branch Guided Swin Transformer delivers superior performance in the challenging task of ischemic

stroke lesion detection and segmentation from brain MRI data. The model's architecture, which combines a hierarchical Swin Transformer backbone with auxiliary pseudo-segmentation supervision, achieves high accuracy, precision, recall, and F1-score demonstrating its effectiveness even in complex imaging conditions. Compared to existing CNN-based and hybrid methods, the proposed model exhibits improved spatial sensitivity and generalization. This advancement aligns with recent findings that vision transformer models, due to their self-attention mechanisms, are capable of capturing long-range dependencies and outperform traditional CNNs in medical imaging tasks [41].

A key factor contributing to this performance is the incorporation of the auxiliary supervision branch, which utilizes bounding box–guided pseudo-masks to regularize intermediate feature representations. This approach promotes spatial coherence during early learning stages, enabling the model to more accurately localize lesions of irregular shape, size, and intensity that is an area where conventional CNNs often fail due to their limited receptive field [42]. Previous works relying on single-task learning or voxel-wise annotation have been constrained by the cost of dense labeling and the inability to generalize to partially labeled datasets. In contrast, our model introduces a weakly supervised strategy that reduces dependency on manual annotations while maintaining segmentation accuracy, echoing the benefits observed in other pseudo-labeling and semi-supervised frameworks [43].

The effectiveness of the Swin Transformer backbone further strengthens the model's adaptability. Its shifted window approach allows for efficient self-attention over both local and global regions without the quadratic computational cost of standard transformers [44]. Recent comparative studies have demonstrated that Swin Transformers are particularly well-suited for medical image analysis, offering a balance between contextual reasoning and computational feasibility [45]. This is especially valuable in stroke diagnosis, where small lesions in deep brain structures or periventricular regions can be easily overlooked by architectures lacking multi-scale attention mechanisms [46]. The stability of convergence curves in Figure 4 and the qualitative alignment of predicted and true lesion regions in Figure 5 both reflect the Swin Transformer's ability to extract consistent spatial hierarchies across image modalities.

Compared to state-of-the-art models summarized in Table 1, the proposed model demonstrates consistent superiority across all key metrics. Several prior approaches using hybrid CNN-ViT architectures reported modest gains in accuracy but lacked robustness in recall or were limited to slice-level predictions [47]. Others, while achieving high precision, did not provide full metric reporting or failed to generalize across datasets with variable contrast and noise characteristics [48]. Our model not only overcomes these limitations but also performs well without relying on full pixel-level annotations highlighting its potential for large-scale deployment in data-constrained clinical environments.

Nonetheless, certain limitations persist. The quality of the pseudo-segmentation masks remains a critical dependency. Noisy or poorly defined bounding boxes can introduce ambiguity into the learning signal, potentially leading to mislocalization or reduced precision [49]. Furthermore, while the Swin Transformer is computationally more efficient than standard ViTs, its inference speed still lags behind lightweight CNNs, which could limit real-time clinical deployment without hardware acceleration [50]. Optimizing the trade-off between model complexity and diagnostic utility remains an important area for future research.

Future directions may involve extending the model to multi-modal learning frameworks by integrating inputs from FLAIR, ADC, and DWI sequences each

offering complementary information about ischemic tissue states [51]. Enhanced pseudo-label generation using consistency training or entropy minimization could further strengthen auxiliary supervision [52]. Additionally, incorporating uncertainty quantification mechanisms could improve clinical interpretability and confidence estimation during deployment [53]. The integration of such tools has shown promise in increasing the clinical acceptability of AI-assisted diagnostic systems [54].

In summary, the proposed Auxiliary Branch Guided Swin Transformer provides a robust and efficient solution for automated stroke lesion detection in brain MRI, leveraging vision transformer principles and weak supervision to address key limitations of existing models. Its ability to generalize across challenging imaging conditions while reducing reliance on dense annotations positions it as a promising direction for future research and clinical translation in neuroimaging diagnostics.

6 CONCLUSION

In this study, we introduced a novel stroke lesion segmentation framework that integrates a Swin Transformer backbone with auxiliary mask supervision, aiming to enhance the accuracy and robustness of lesion detection in brain MRI. The proposed model leverages the hierarchical attention mechanism of the Swin Transformer to capture both local and global contextual information, while the auxiliary branch utilizes pseudo segmentation masks to provide intermediate spatial supervision during training. Experimental results on the ISLES 2024 dataset demonstrate that our approach outperforms several state-of-the-art CNN- and transformer-based models in terms of Dice similarity, precision, and recall. The qualitative and quantitative analyses further confirm the model's ability to accurately localize and delineate stroke lesions of varying size and complexity. The auxiliary supervision component significantly improves feature learning and model generalization, especially in challenging cases with small or diffusely distributed lesions. Although limitations related to computational cost and dependency on pseudo-mask quality remain, the proposed architecture presents a scalable and effective solution for automated stroke assessment. Future work will focus on enhancing the model through multi-modal data integration, domain adaptation, and uncertainty modeling. Overall, this research contributes to the advancement of transformer-based segmentation approaches in neuroimaging and supports their clinical applicability in stroke diagnosis.

7 ACKNOWLEDGEMENTS

This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23489899 "Applying Deep Learning and Neuroimaging Methods for Brain Stroke Diagnosis."

8 REFERENCES

[1] P. B. Nielsen, R. F. Brøndum, A. K. Nøhr, T. F. Overvad, and G. Y. H. Lip, "Risk of stroke in male and female patients with atrial fibrillation in a nationwide cohort," *Nature Communications*, vol. 15, no. 1, p. 6728, 2024. https://doi.org/10.1038/s41467-024-51193-0

[2] R. Geetha, E. Priya, and M. Vijayakumar, "An approach for automated acute cerebral ischemic stroke lesion segmentation and correlation of significant features with modified Rankin Scale," *Biomedical Signal Processing and Control*, vol. 100, p. 106921, 2025. https://doi.org/10.1016/j.bspc.2024.106921

[3] A. Ydyrys, L. Sarybekova, and N. Tleukhanova, "The multipliers of multiple trigonometric Fourier series," *Open Engineering*, vol. 6, no. 1, 2016. https://doi.org/10.1515/eng-2016-0046

[4] S. U. R. Khan, S. Asif, M. Zhao, W. Zou, Y. Li, and X. Li, "Optimized deep learning model for comprehensive medical image analysis across multiple modalities," *Neurocomputing*, vol. 619, p. 129182, 2024. https://doi.org/10.1016/j.neucom.2024.129182

[5] C. C. Ukwuoma *et al.*, "Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable AI – LIME & SHAP," *Biomedical Signal Processing and Control*, vol. 100, p. 107014, 2024. https://doi.org/10.1016/j.bspc.2024.107014

[6] C. Chen, N. A. M. Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," *Computers in Biology and Medicine*, vol. 185, p. 109507, 2024. https://doi.org/10.1016/j.compbiomed.2024.109507

[7] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, "Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases," *Journal of Imaging Informatics in Medicine*, 2025. https://doi.org/10.1007/s10278-025-01481-y

[8] R. Mousa, B. Rezaei, L. Mahmoudi, and J. Abdollahi, "Multi-modal wound classification using wound image and location by Swin Transformer and transformer," *Expert Systems with Applications*, vol. 280, p. 127077, 2025. https://doi.org/10.1016/j.eswa.2025.127077

[9] D. D. Himabindu, E. L. Lydia, M. V. Rajesh, M. A. Ahmed, and M. K. Ishak, "Leveraging swin transformer with ensemble of deep learning model for cervical cancer screening using colposcopy images," *Scientific Reports*, vol. 15, no. 1, p. 7900, 2025. https://doi.org/10.1038/s41598-025-90415-3

[10] H. Wu, F. Xiao, and C. Liang, "Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation," in *Computer Vision – ECCV 2022*, in Lecture Notes in Computer Science, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13680, 2022, pp. 417–434. https://doi.org/10.1007/978-3-031-20044-1_24

[11] M. T. R. Shawon, G. M. S. Shibli, F. Ahmed, and S. K. S. Joy, "Explainable cost-sensitive deep neural networks for brain tumor detection from brain MRI images considering data imbalance," *Multimedia Tools and Applications*, vol. 84, pp. 43615–43642, 2025. https://doi.org/10.1007/s11042-025-20842-x

[12] A. Tursynova, B. Omarov, A. Sakhipov, and N. Tukenova, "Brain stroke lesion segmentation using computed Tomography Images based on modified U-Net model with ResNet blocks," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 13, pp. 97–112, 2022. https://doi.org/10.3991/ijoe.v18i13.32881

[13] T. K. Dutta, D. R. Nayak, and Y. Zhang, "ARM-Net: Attention-guided residual multiscale CNN for multiclass brain tumor classification using MR images," *Biomedical Signal Processing and Control*, vol. 87, pp. 105421–105421, 2024. https://doi.org/10.1016/j.bspc.2023.105421

[14] S. Raminedi, S. Shridevi, and D. Won, "Multi-modal transformer architecture for medical image analysis and automated report generation," *Scientific Reports*, vol. 14, no. 1, p. 19281, 2024. https://doi.org/10.1038/s41598-024-69981-5

[15] S. Takahashi *et al.*, "Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review," *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 2024. https://doi.org/10.1007/s10916-024-02105-8

[16] J. W. Kim, A. U. Khan, and I. Banerjee, "Systematic review of hybrid vision transformer architectures for radiological image analysis," *Journal of Imaging Informatics in Medicine*, vol. 38, pp. 3248–3262, 2025. https://doi.org/10.1007/s10278-024-01322-4

[17] F. Ghazouani, P. Véra, and S. Ruan, "Efficient brain tumor segmentation using Swin transformer and enhanced local self-attention," *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, pp. 273–281, 2023. https://doi.org/10.1007/s11548-023-03024-8

[18] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From CNN to transformer: A review of medical image segmentation models," *Journal of Imaging Informatics in Medicine*, vol. 37, pp. 1529–1547, 2024. https://doi.org/10.1007/s10278-024-009817

[19] H. Yang and D. Yang, "CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images," *Expert Systems with Applications*, vol. 213, p. 119024, 2023. https://doi.org/10.1016/j.eswa.2022.119024

[20] Z. Ning, S. Zhong, Q. Feng, W. Chen, and Y. Zhang, "SMU-Net: Saliency-guided morphology-aware U-Net for breast lesion segmentation in ultrasound image," *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 476–490, 2022. https://doi.org/10.1109/TMI.2021.3116087

[21] H. Wu, F. Xiao, and C. Liang, "Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation," in *Computer Vision – ECCV 2022*, in Lecture Notes in Computer Science, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13680, 2022, pp. 417–434. https://doi.org/10.1007/978-3-031-20044-1_24

[22] M. Sadeghibakhi, H. Pourreza, and H. Mahyar, "Multiple sclerosis lesions segmentation using attention-based CNNs in FLAIR images," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–11, 2022. https://doi.org/10.1109/JTEHM.2022.3172025

[23] M. Bento, I. Fantini, J. Park, L. Rittner, and R. Frayne, "Deep learning in large and multi-site structural brain MR imaging datasets," *Frontiers in Neuroinformatics*, vol. 15, p. 805669, 2022. https://doi.org/10.3389/fninf.2021.805669

[24] A. J. Priana, H. Tolle, I. Aknuranda, and E. Aristijono, "User experience design of stroke patient communications using Mobile Finger (MOFI) communication board with user center design approach," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 12, no. 2, pp. 162–176, 2018. https://doi.org/10.3991/ijim.v12i2.7937

[25] C. Sendra-Balcells *et al.*, "Domain generalization in deep learning for contrast-enhanced imaging," *Computers in Biology and Medicine*, vol. 149, p. 106052, 2022. https://doi.org/10.1016/j.compbiomed.2022.106052

[26] Q. Qian, Y. Wang, T. Zhang, and Y. Qin, "Maximum mean square discrepancy: A new discrepancy representation metric for mechanical fault transfer diagnosis," *Knowledge-Based Systems*, vol. 276, p. 110748, 2023. https://doi.org/10.1016/j.knosys.2023.110748

[27] R. Seah, H. Zhou, M. Jalaleddine, and W. J. Gross, "xSA: A binary cross-entropy simulated annealing polar decoder," in *2023 12th International Symposium on Topics in Coding (ISTC)*, 2023, pp. 1–5. https://doi.org/10.1109/ISTC57237.2023.10273491

[28] Y. Wu *et al.*, "Enhancing Diffusion-Weighted Images (DWI) for diffusion MRI: Is it enough without non-diffusion-weighted B = 0 reference?" in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 2025, pp. 1–4. https://doi.org/10.1109/ISBI60581.2025.10980956

[29] E. O. Riedel *et al.*, "ISLES 2024: The first longitudinal multimodal multi-center real-world dataset in (sub-)acute stroke," *arXiv preprint arXiv:2408.11142*, 2024. https://doi.org/10.48550/arxiv.2408.11142

[30] S. J. Ahn, T. Taoka, W. Moon, and S. Naganawa, "Contrast-enhanced fluid-attenuated inversion recovery in neuroimaging: A narrative review on clinical applications and technical advances," *Journal of Magnetic Resonance Imaging*, vol. 56, no. 2, pp. 341–353, 2022. https://doi.org/10.1002/jmri.28117

[31] Yì Xiáng, J. Wáng, K.-X. Zhao, F.-Z. Ma, and B.-H. Xiao, "The contribution of T2 relaxation time to MRI-derived apparent diffusion coefficient (ADC) quantification and its potential clinical implications," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 10, pp. 7410–7416, 2023. https://doi.org/10.21037/qims-23-1106

[32] Y. M. Wong *et al.*, "Machine learning prediction of Dice similarity coefficient for validation of deformable image registration," *Intelligence-Based Medicine*, vol. 10, p. 100163, 2024. https://doi.org/10.1016/j.ibmed.2024.100163

[33] D. Black, W. Li, and S. Molloi, "Optimized Hausdorff distance loss function based on a GPU-Accelerated distance transform," *TechRxiv*, 2023. https://doi.org/10.36227/techrxiv.23612100.v1

[34] W. Cullerne Bown, "Sensitivity and specificity versus precision and recall, and related dilemmas," *Journal of Classification*, vol. 41, pp. 402–426, 2024. https://doi.org/10.1007/s00357-024-09478-y

[35] R. Raj, J. Mathew, S. K. Kannath, and J. Rajan, "StrokeViT with AutoML for brain stroke classification," *Engineering Applications of Artificial Intelligence*, vol. 119, p. 105772, 2023. https://doi.org/10.1016/j.engappai.2022.105772

[36] A. Tursynova *et al.*, "Deep learning-enabled brain stroke classification on computed tomography images," *Computers, Materials & Continua*, vol. 75, no. 1, pp. 1431–1446, 2023. https://doi.org/10.32604/cmc.2023.034400

[37] W. Abbaoui, S. Retal, S. Ziti, B. E. Bhiri, and H. Moussif, "Ischemic stroke classification using VGG-16 convolutional neural networks: A study on Moroccan MRI scans," *International Journal of Online and Biomedical Engineering*, vol. 20, no. 2, pp. 61–77, 2024. https://doi.org/10.3991/ijoe.v20i02.44845

[38] H. Yu *et al.*, "Prognosis of ischemic stroke predicted by machine learning based on multimodal MRI radiomics," *Frontiers in Psychiatry*, vol. 13, 2023. https://doi.org/10.3389/fpsyt.2022.1105496

[39] A. B. Sreekumari and A. T. Y. Paulsy, "Hybrid deep learning based stroke detection using CT images with routing in an IoT environment," *Network Computation in Neural Systems*, pp. 1–40, 2025. https://doi.org/10.1080/0954898X.2025.2452280

[40] R. Qasrawi *et al.*, "Hybrid ensemble deep learning model for advancing ischemic brain stroke detection and classification in clinical application," *Journal of Imaging*, vol. 10, no. 7, p. 160, 2024. https://doi.org/10.3390/jimaging10070160

[41] M. Nouman, M. Mabrok, and E. A. Rashed, "Neuro-TransUNet: Segmentation of stroke lesion in MRI using transformers," *arXiv preprint arXiv:2406.06017*, 2024. https://doi.org/10.48550/arXiv.2406.06017

[42] W. K. Soh and J. C. Rajapakse, "Hybrid UNet and transformer network for ischemic stroke segmentation using MRI and CT datasets," *Frontiers in Neuroscience*, vol. 17, p. 1298514, 2023. https://doi.org/10.3389/fnins.2023.1298514

[43] C. Foulon *et al.*, "Generalizable automated ischaemic stroke lesion segmentation with vision transformers," *arXiv preprint arXiv:2502.06939*, 2025. https://doi.org/10.48550/arXiv.2502.06939

[44] R. Ahmed, A. Al Shehhi, N. Werghi, and M. L. Seghier, "Segmentation of stroke lesions using transformers augmented MRI analysis," *Human Brain Mapping*, vol. 45, no. 11, p. e26803, 2024. https://doi.org/10.1002/hbm.26803

[45] Q. Cao and X. Cheng, "A hybrid feature fusion deep learning framework for multi-source medical image analysis," *Information Processing & Management*, vol. 62, no. 1, p. 103934, 2024. https://doi.org/10.1016/j.ipm.2024.103934

[46] B. P. Garcia-Salgado *et al.*, "Enhanced ischemic stroke lesion segmentation in MRI using Attention U Net with generalized dice focal loss," *Applied Sciences*, vol. 14, no. 18, p. 8183, 2024. https://doi.org/10.3390/app14188183

[47] W. K. Soh and J. C. Rajapakse, "Noise induced self supervised hybrid UNet transformer for ischemic stroke segmentation," *Scientific Reports*, vol. 15, no. 1, p. 19783, 2025. https://doi.org/10.1038/s41598-025-04819-2

[48] Y. Zhang, W. Wu, H. Chen, and J. Huang, "Cross-modality medical image segmentation using a dual-stream hybrid Transformer," *Medical Image Analysis*, vol. 89, p. 102905, 2024. https://doi.org/10.1016/j.media.2023.102905

[49] X. Zhao and W. Wang, "Semi supervised medical image segmentation based on deep consistent collaborative learning," *Journal of Imaging*, vol. 10, no. 5, p. 118, 2024. https://doi.org/10.3390/jimaging10050118

[50] Q. Pu, Z. Xi, S. Yin, Z. Zhao, and L. Zhao, "Advantages of transformer and its application for medical image segmentation: A survey," *BioMedical Engineering Online*, vol. 23, no. 1, p. 14, 2024. https://doi.org/10.1186/s12938-024-01212-4

[51] L. Tomasetti *et al.*, "Self supervised few shot learning for ischemic stroke lesion segmentation," *arXiv preprint arXiv:2303.01332*, 2023. https://doi.org/10.48550/arXiv.2303.01332

[52] W.-S. Ryu *et al.*, "Deep learning-based automatic segmentation of cerebral infarcts on diffusion MRI," *Scientific Reports*, vol. 15, no. 1, p. 13214, 2025. https://doi.org/10.1038/s41598-025-91032-w

[53] M. Nouman, M. Mabrok, and E. A. Rashed, "Neuro-TransUNet: U-Net with SwinUNETR fusion for MRI stroke lesion segmentation," *arXiv preprint arXiv:2406.06017*, 2024. https://doi.org/10.48550/arXiv.2406.06017

[54] D. Wang, Z. Wang, L. Chen, H. Xiao, and B. Yang, "Cross-parallel transformer: Parallel VIT for medical image segmentation," *Sensors*, vol. 23, no. 23, p. 9488, 2023. https://doi.org/10.3390/s23239488

9 AUTHORS

Batyrkhan Omarov received his bachelor's and master's degrees from Al-Farabi Kazakh National University, Almaty, Kazakhstan in 2008 and 2010, respectively. In 2019, he received his Ph.D. from Tenaga National University, Kuala Lumpur, Malaysia. His research interests include machine learning, natural language processing, robotics, and artificial intelligence in medicine (E-mail: batyahan@gmail.com).

Zhanseri Ikram is a PhD student at Al-Farabi Kazakh National University, Almaty, Kazakhstan. His research interests are machine learning, deep learning, image processing, and computer vision (E-mail: zhanserikz@gmail.com).