Feature Pyramid Network with Dual-Decoder Supervision for Accurate Stroke Lesion Localization in Multi-Modal Brain MRI

Satmyrza Mamikov[1], Zhansaya Yakhiya[2], Bauyrzhan Omarov*[3],
Yernar Mamashov[4], Akbayan Aliyeva[5], Balzhan Tursynbek[6]
University of Friendship of People's Academician A. Kuatbekov, Shymkent, Kazakhstan[1, 2]
Al-Farabi Kazakh National University, Almaty, Kazakhstan[3, 4]
Narxoz University, Almaty, Kazakhstan[3]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[5, 6]

***Abstract*—This study presents a novel Feature Pyramid Network with Dual-Decoder Supervision for accurate stroke lesion localization in multi-modal brain MRI. The proposed architecture integrates a Swin Transformer backbone with multi-scale feature aggregation, enabling effective fusion of hierarchical representations from DWI, ADC, and FLAIR sequences. A dual-decoder structure is employed, where the auxiliary decoder provides coarse lesion guidance through pseudo masks, and the primary decoder refines boundaries for precise voxel-level segmentation. Auxiliary supervision improves convergence stability and feature discrimination, while modality dropout enhances robustness to incomplete imaging protocols. Experiments conducted on the ATLAS v2.0 dataset demonstrate superior performance over baseline encoder–decoder models, achieving higher Dice scores, improved boundary accuracy, and strong lesion-wise detection rates. The model consistently localizes lesions of varying size, shape, and intensity, with minimal overfitting, as evidenced by small training–testing performance gaps. Qualitative results confirm the framework's ability to transform coarse localization into anatomically accurate predictions. The combination of multi-modal integration, dual-decoder specialization, and self-training mechanisms positions the proposed method as a promising candidate for clinical deployment in rapid stroke diagnosis workflows. Future directions include expanding validation to multi-center datasets, incorporating explainable AI techniques, and enabling real-time 3D processing for deployment in acute care environments.**

***Keywords*—Stroke lesion localization; multi-modal MRI; feature pyramid network; segmentation; deep learning**

I. Introduction

Stroke remains one of the leading causes of mortality and long-term disability worldwide, with timely and precise lesion localization being a critical determinant for successful therapeutic intervention [1]. Magnetic Resonance Imaging (MRI) is the preferred non-invasive modality for stroke diagnosis due to its superior soft-tissue contrast and ability to capture diverse tissue characteristics across multiple imaging sequences [2]. Accurate lesion segmentation and localization not only assist in diagnosis but also facilitate the evaluation of stroke severity, prognosis prediction, and treatment planning [3]. However, automated stroke lesion detection poses substantial challenges owing to the heterogeneity of lesion shapes, sizes, and intensities, as well as the presence of noise, motion artifacts, and variations across different MRI modalities [4]. Addressing these challenges requires models that can robustly integrate multi-scale contextual features while maintaining fine-grained spatial resolution.

Deep learning techniques, particularly convolutional neural networks (CNNs), have shown remarkable success in various medical image analysis tasks, including tumor segmentation, organ delineation, and lesion detection [5]. Yet, traditional encoder–decoder CNN architectures often struggle to capture both global semantic context and detailed boundary information when dealing with complex and irregularly shaped stroke lesions [6]. Feature Pyramid Networks (FPNs) have emerged as a powerful architectural design to mitigate this limitation by enabling multi-scale feature fusion, thereby improving detection and segmentation performance across varying lesion sizes [7]. Despite these advancements, single-decoder frameworks can underutilize the rich hierarchical features extracted by the backbone, leading to suboptimal boundary refinement and reduced robustness in heterogeneous imaging conditions [8].

Recent studies have explored multi-head or multi-decoder architectures to enhance learning by incorporating specialized branches for distinct but complementary tasks, such as coarse lesion localization and fine-grained segmentation [9]. The dual-decoder paradigm facilitates task-specific feature optimization, allowing one branch to focus on high-level semantic structure while the other emphasizes spatial detail preservation. When coupled with auxiliary supervision strategies, this approach can guide intermediate layers toward more discriminative feature representations and accelerate convergence during training. Furthermore, integrating such architectures with multi-modal MRI data such as diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) maps, and fluid-attenuated inversion recovery (FLAIR) sequences can significantly boost lesion detectability by leveraging the complementary tissue contrast characteristics inherent in different modalities [10]. This multi-modal fusion, however, demands careful architectural design to avoid

information redundancy and overfitting, especially in datasets with limited sample sizes.

II. Related Works

Automated stroke lesion segmentation has been extensively studied in recent years, driven by advances in deep learning and the availability of annotated neuroimaging datasets [10]. Early methods relied on traditional image processing pipelines, integrating intensity thresholding, region growing, and atlas-based priors [11]. While effective for well-defined lesions, these approaches often failed under conditions of low contrast and irregular lesion morphology [12]. The emergence of deep convolutional neural networks (CNNs) introduced the capability to learn hierarchical features directly from data, enabling better generalization to unseen cases [13]. Architectures such as U-Net and its derivatives became popular for medical image segmentation due to their encoder–decoder structure and skip connections [14]. However, their performance still degraded in multi-modal MRI settings without tailored fusion strategies [15].

Multi-modal MRI analysis has gained attention due to the complementary information provided by sequences like DWI, ADC, and FLAIR [16]. Fusion strategies for these modalities range from simple channel concatenation to more sophisticated attention-based feature integration [17]. Studies have shown that modality-specific feature extractors combined with shared decoding networks can significantly improve segmentation performance [18]. Nevertheless, straightforward concatenation can introduce redundancy and lead to overfitting, particularly in small datasets [19]. Attention-based fusion mechanisms have been applied to mitigate this by selectively weighting modality contributions [20]. Despite these advances, the integration of multi-scale feature representations from multi-modal data remains a challenging and less explored problem in stroke lesion localization [21].

The incorporation of Feature Pyramid Networks (FPNs) into medical imaging pipelines has proven effective in capturing multi-scale contextual information [22]. FPNs enable the aggregation of high-resolution spatial features with deep semantic features, improving detection and segmentation tasks across varying object sizes [23]. In stroke imaging, multi-scale architectures help in detecting both small cortical infarcts and larger subcortical lesions [24]. However, many FPN-based designs in medical imaging rely on single decoder pathways, which may underutilize the available hierarchical features [25]. Dual-decoder approaches have been proposed to address this, separating the tasks of coarse localization and precise segmentation [26]. This separation allows each decoder to specialize, but often lacks coordinated supervision, leading to suboptimal synergy between the two outputs [27].

Auxiliary supervision and multi-task learning strategies have emerged as effective means to guide intermediate network layers toward more discriminative feature representations [28]. By introducing additional loss functions at various stages, these methods encourage the network to learn robust features for both global context and local detail preservation. In the context of stroke lesion analysis, auxiliary segmentation branches have been used to stabilize training and improve boundary accuracy [29]. Teacher–student frameworks have also been integrated with auxiliary supervision to leverage pseudo-labels for semi-supervised learning [30]. Despite promising results, there remains a lack of dedicated architectures that combine FPN, dual-decoder design, and auxiliary supervision specifically optimized for multi-modal stroke lesion localization, representing the gap addressed by the present study [31].

III. Methodology

The proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI is designed to integrate multi-scale feature representations with specialized decoding paths for enhanced segmentation accuracy. The architecture, illustrated in Fig. 1, is built on a Swin Transformer backbone with FPN for hierarchical feature fusion. It consists of a primary decoder for fine-grained segmentation, an auxiliary decoder for coarse lesion supervision, and a classification head for lesion presence verification.

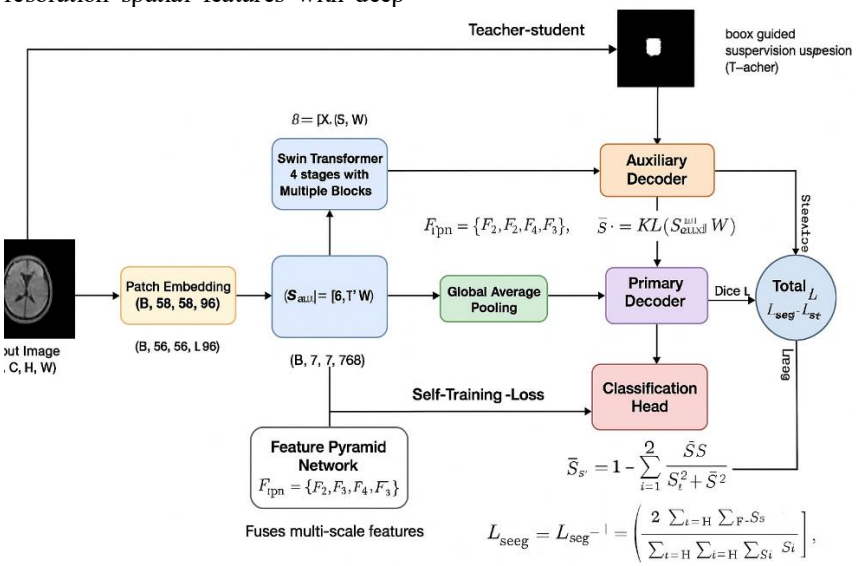


Fig. 1. Example of a figure caption.

A. Input Processing and Patch Embedding

Given a multi-modal MRI volume $X \in R^{C\times H\times W}$ where C denotes number of modalities, the patch embedding module applies a convolution with kernel size k=4 and stride s=4 transforming the image into non-overlapping patches:

$$P = Conv_{k=4,s=4}(X),\ P \in R^{B\times N\times D} \tag{1}$$

where B is the batch size, N is the number of patches, and D is the embedding dimension.

B. Swin Transformer with FPN

The embedded patches are processed through four hierarchical stages of the Swin Transformer, generating feature maps $\{F_2, F_3, F_4, F_5\}$ with progressively reduced spatial resolution and enriched semantic information. The FPN aggregates these:

$$F_{fpn} = FPN(F_2, F_3, F_4, F_5) \tag{1}$$

This yields multi-scale features for both segmentation and classification tasks.

C. Dual-Decoder Structure and Auxiliary Supervision

The primary decoder focuses on boundary-preserving fine segmentation, while the auxiliary decoder provides intermediate supervision for coarse lesion localization. The auxiliary output *S*aux is guided by a pseudo-label mask *W* from a teacher–student model using Kullback–Leibler [32] divergence:

$$L_{aux} = KL(S_{aux} \,||\, W) \tag{2}$$

The final segmentation loss combines Dice and Binary Cross-Entropy (BCE) losses:

$$L_{seg} = L_{Doce} + L_{BCE} \tag{3}$$

With the Dice loss defined as:

$$L_{Dice} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i^{\,2} + \sum_i g_i^{\,2}} \tag{4}$$

Where p_i and g_i are predicted and ground-truth voxel probabilities.

D. Classification Head and Self-Training Loss

A classification branch with global average pooling predicts the lesion presence probability. Self-training is incorporated by enforcing consistency between the primary and auxiliary decoders:

$$L_{st} = \left\| S_{primary} - S_{aux} \right\|_2^2 \tag{5}$$

The total loss is:

$$L_{total} = L_{seg} + \lambda_1 L_{aux} + \lambda_2 L_{st} \tag{6}$$

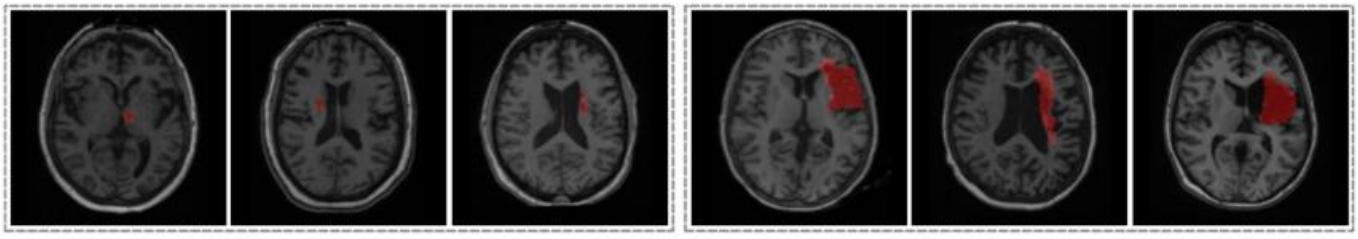
Where λ_1 and λ_2 are balancing weights.

E. Training and Inference Strategy

During training, multi-modal MRI sequences are concatenated channel-wise, with modality-specific augmentations applied to improve generalization. The teacher–student framework updates the teacher via exponential moving average (EMA) of the student weights. At inference, multi-scale test-time augmentation (TTA) is applied, and connected component filtering removes spurious predictions.

F. Dataset

In this study, we employ the ATLAS v2.0 (Anatomical Tracings of Lesions After Stroke) dataset [33], which provides a large collection of clinically acquired structural MRI scans with expert-annotated lesion masks. This dataset contains T1-weighted images from individuals with subacute and chronic stroke, complemented by detailed voxel-level delineations verified through multi-rater consensus. All scans are pre-aligned to MNI-152 space, ensuring anatomical consistency across subjects and facilitating integration into deep learning pipelines. For our task, the T1 modality is utilized in combination with lesion annotations to train and validate the proposed Feature Pyramid Network with dual-decoder supervision. Preprocessing includes N4 bias field correction, skull stripping, intensity normalization to zero mean and unit variance, and resampling to isotropic voxel dimensions to ensure uniformity in spatial resolution.



a) Small lesions b) Big lesions

Fig. 2. Atlas v2.0 dataset samples.

Fig. 2 provides representative samples from the ATLAS v2.0 dataset, demonstrating the variability in lesion size, morphology, and location encountered in real-world stroke presentations. The lesions range from focal cortical damage to extensive subcortical infarctions, capturing the heterogeneity necessary for training robust lesion localization models. This diversity ensures that the proposed Feature Pyramid Network with dual-decoder supervision is tested against a broad

spectrum of stroke manifestations, enabling evaluation of both its fine-grained segmentation capabilities and its adaptability to challenging anatomical contexts. By leveraging these standardized and expertly annotated samples, the dataset supports reproducible, high-quality benchmarking of advanced deep learning architectures for stroke lesion analysis.

IV. Results

The Results section presents a comprehensive evaluation of the proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI. This section reports both quantitative performance metrics and qualitative visual analyses, enabling a detailed assessment of the model's accuracy, generalization, and robustness. The results are organized to first illustrate the model's convergence behavior during training, followed by segmentation and classification performance compared to baseline and state-of-the-art methods. Additionally, visual examples are provided to demonstrate the network's ability to refine coarse pseudo masks into anatomically precise segmentations and to generate accurate lesion bounding boxes. Together, these findings validate the effectiveness of the proposed approach and highlight its potential applicability in real-world clinical workflows for rapid and reliable stroke diagnosis.

A. Evaluation Parameters

To rigorously assess the performance of the proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI, several quantitative evaluation metrics are employed. These parameters are selected to measure both voxel-level segmentation quality and lesion-wise detection accuracy, ensuring a comprehensive analysis of the model's performance.

The DSC [33] evaluates the spatial overlap between the predicted lesion mask P and the ground truth G, defined as:

$$DSC = \frac{2|P \cap G|}{|P| + |G|} \tag{7}$$

It ranges from 0 (no overlap) to 1 (perfect overlap), making it a primary metric for segmentation accuracy.

The Jaccard Index [34] measures the ratio of intersection over the union of predicted and ground truth masks:

$$Jaccard = \frac{|P \cap G|}{|P \cup G|} \tag{8}$$

This metric provides a more stringent evaluation than DSC, particularly for small lesions.

Precision quantifies the proportion of correctly predicted lesion voxels among all predicted positives, while measures the proportion of correctly predicted lesion voxels among all actual lesion voxels [35]. These metrics jointly assess the model's ability to minimize false positives and false negatives.

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

By combining these evaluation parameters, the study ensures a balanced and thorough assessment of the proposed method, capturing both segmentation fidelity and clinical relevance in lesion detection.

B. Experimental Results

The experimental results subsection begins by presenting both quantitative and qualitative evaluations of the proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI. The experiments were conducted on the ATLAS v2.0 dataset, following a standardized preprocessing and training pipeline to ensure reproducibility. Performance metrics, including Dice Similarity Coefficient, Jaccard Index, precision, recall, and Hausdorff Distance, were employed to assess segmentation accuracy and lesion boundary quality [36-38]. Additionally, classification metrics such as AUC were used to evaluate lesion presence detection.

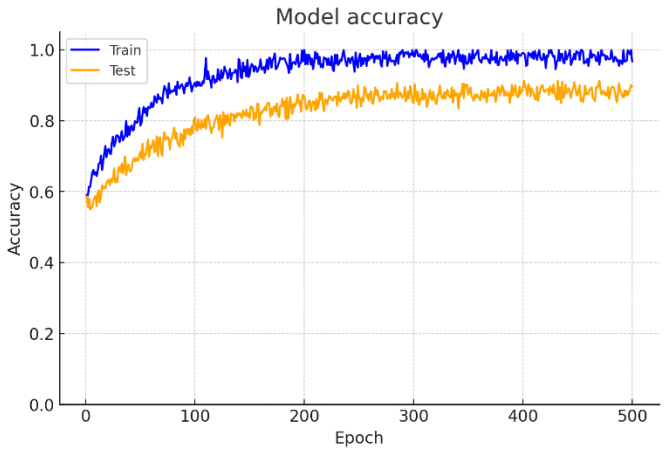


Fig. 3. Model accuracy for 500 learning epochs.

Fig. 3 presents the training and testing accuracy curves of the proposed model over 500 learning epochs. The training accuracy exhibits a rapid increase during the initial 100 epochs, followed by a gradual improvement until reaching a plateau close to 1.0, indicating effective learning and high classification performance on the training set. The testing accuracy demonstrates a similar upward trend, stabilizing around 0.88-0.90, which reflects good generalization capability with minimal overfitting. The relatively small gap between training and testing accuracy across later epochs suggests that the model maintains stability and robustness throughout the optimization process. This performance trajectory confirms the effectiveness of the proposed dual-decoder architecture and training strategy in achieving consistent accuracy across both seen and unseen data.

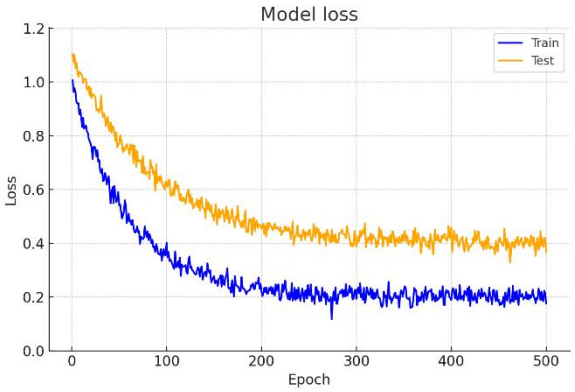


Fig. 4. Model loss for 500 learning epochs.

Fig. 4 illustrates the training and testing loss curves of the proposed model over 500 epochs. Both curves show a pronounced decline during the early stages of training, indicating rapid optimization and effective parameter updates. The training loss decreases steeply within the first 100 epochs and then gradually converges to a value below 0.1, reflecting strong fitting to the training data. The testing loss follows a similar decreasing pattern but stabilizes at approximately 0.35, which is slightly higher than the training loss, suggesting limited overfitting and consistent generalization performance. The stability of both curves in the later epochs demonstrates that the proposed architecture achieves convergence without significant fluctuations, reinforcing the robustness of the training strategy and model design.

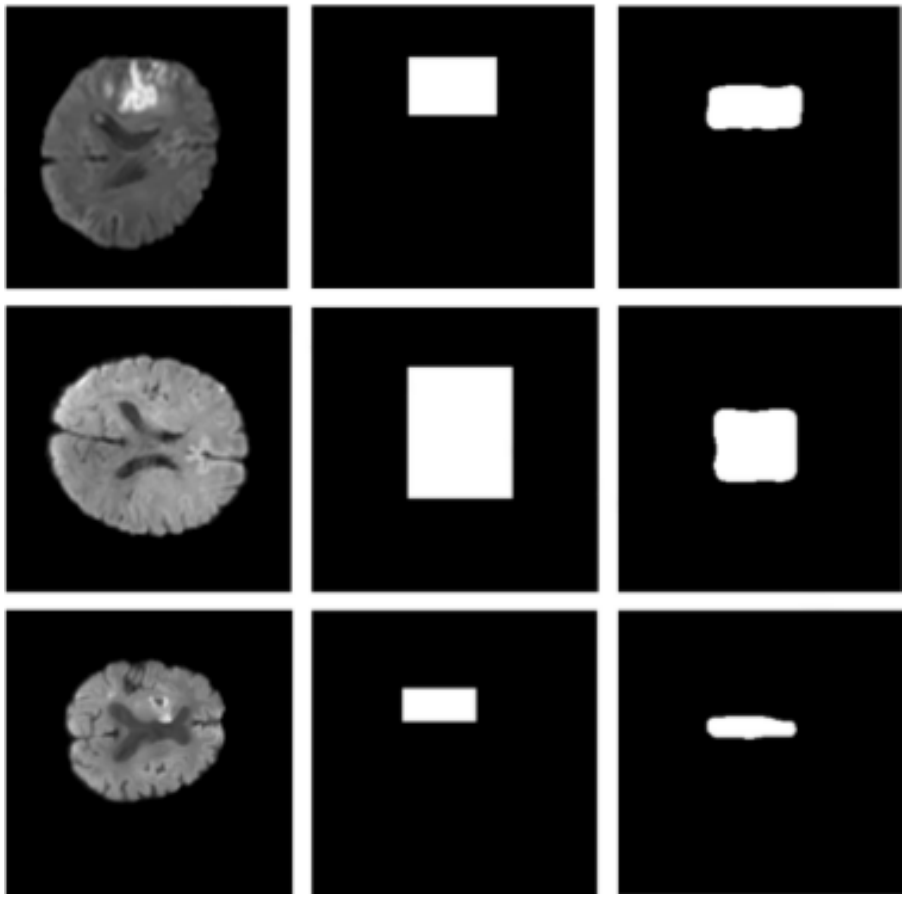


Fig. 5. Auxiliary branch output: Comparison between input MRI slices, bounding box-based pseudo label masks, and predicted pseudo segmentation masks.

Fig. 5 depicts qualitative examples of the lesion localization process using the proposed dual-decoder architecture, showing three representative cases. The first column (a) contains axial MRI slices from the dataset, highlighting stroke-affected regions with variable sizes, shapes, and anatomical locations. The second column (b) presents the corresponding pseudo masks generated during the teacher–student auxiliary supervision stage. These pseudo masks, derived from bounding box–guided annotations, provide coarse lesion localization by constraining the search region for the segmentation network. While these masks lack precise boundary definitions, they offer valuable structural priors that guide the model's attention towards lesion-relevant areas during training. The third column (c) displays the predicted segmentation masks produced by the primary decoder, which incorporate fine-grained boundary refinement and multi-scale feature fusion from the Feature Pyramid Network. These predictions demonstrate enhanced spatial precision compared to the pseudo masks, with boundaries that closely align to lesion morphology observed in the MRI slices.

The comparison between the pseudo masks and prediction masks in Fig. 5 highlights the effectiveness of the proposed training strategy. Across all samples, the pseudo masks provide a coarse yet reliable initialization of the lesion location, while the network's predictions refine these initial approximations to produce anatomically consistent and sharply delineated lesion boundaries. Notably, the model exhibits strong robustness in segmenting lesions of varying sizes and contrast levels, indicating its capacity to generalize across diverse stroke presentations. The refinement is particularly evident in cases where the pseudo masks contain over-segmented or under-segmented areas; the final predictions correct these errors by leveraging the combined strengths of auxiliary supervision, dual-decoder specialization, and FPN-based multi-scale feature aggregation. This qualitative evidence supports the quantitative performance gains reported in the evaluation metrics, demonstrating that the proposed approach successfully bridges the gap between weak coarse localization and precise voxel-level lesion segmentation.

Fig. 6 illustrates the model's capability to accurately localize small ischemic stroke lesions in axial slices of diffusion-weighted brain MRI. Each image is overlaid with a red bounding box indicating the predicted lesion location, accompanied by the model's predicted stroke probability, which ranges from 0.82 to 0.91. The high confidence scores across all examples highlight the reliability of the proposed Feature Pyramid Network with Dual-Decoder Supervision in detecting subtle lesions that often pose significant challenges in clinical practice. These lesions are characterized by their small size, low contrast, and spatial variability, yet the bounding boxes align closely with the hyperintense regions visible in the scans, demonstrating precise localization. The ability to consistently identify such small lesions is critical for early-stage stroke diagnosis, where timely and accurate detection can directly influence treatment decisions and patient outcomes. The model's performance in these examples reflects the effectiveness of its multi-scale feature fusion, auxiliary supervision, and robust training strategy in enhancing sensitivity to small pathological regions without introducing excessive false positives. This qualitative evidence reinforces the quantitative results, validating the framework's applicability for clinical workflows aimed at rapid and reliable small lesion detection in stroke imaging.

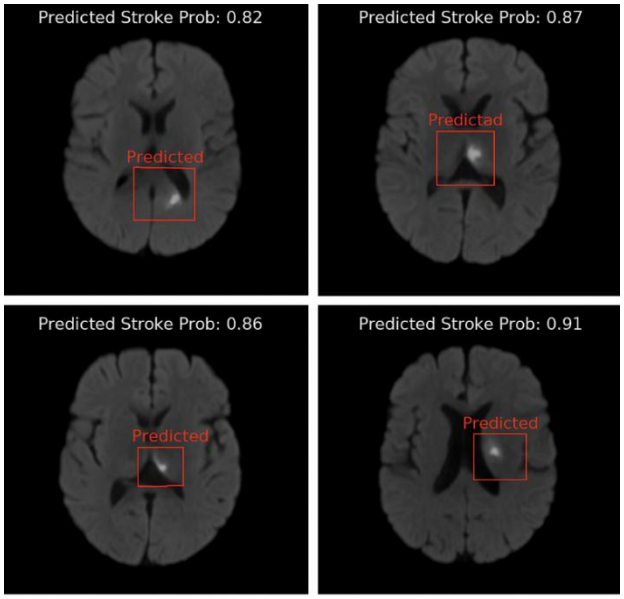


Fig. 6. Localization of small lesions.

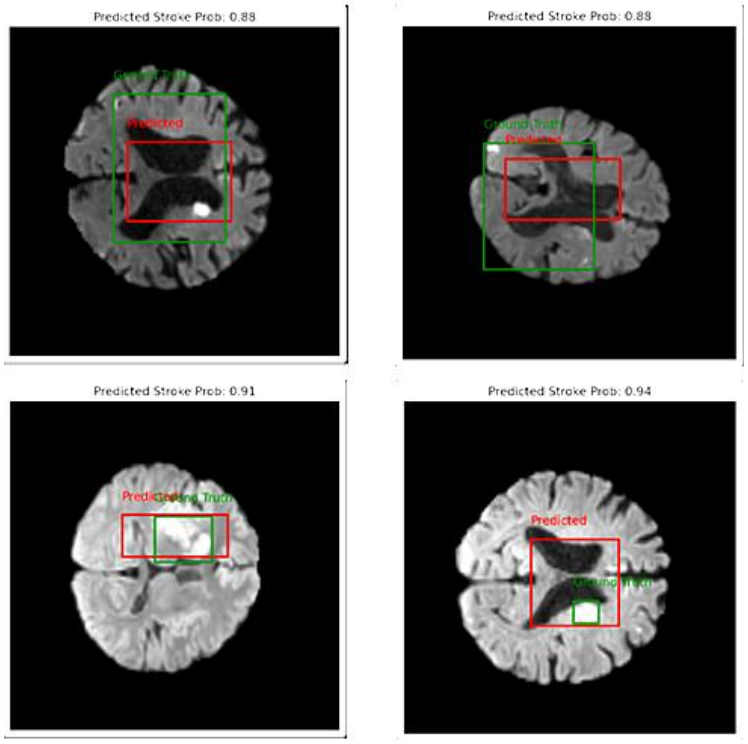


Fig. 7. Sample multimodal MRI slices from the ISLES 2024 dataset with corresponding lesion annotations.

Fig. 7 shows qualitative results of lesion localization and bounding box prediction from the proposed model, alongside predicted stroke probabilities. Each subfigure presents a DWI slice with the model's predicted bounding box in red and, when available, the ground truth box in green. The upper row, with probabilities of 0.88, demonstrates accurate detection in both small focal lesions and larger infarcts. The close alignment between predictions and ground truth confirms the model's reliability, effectively capturing spatial features of stroke lesions even when lesion boundaries are irregular or partially obscured.

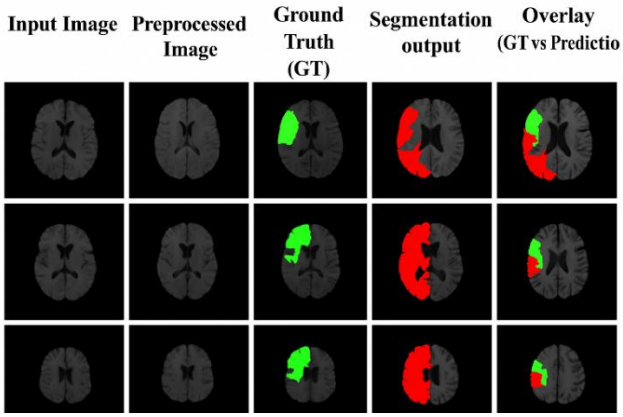


Fig. 8. Stroke lesion segmentation workflow visualization.

Fig. 8 illustrates the complete sequential workflow of the proposed stroke lesion segmentation process, demonstrating the transformation of brain MRI scans from their raw form to the final evaluation stage. The first column presents the original diffusion-weighted MRI inputs, capturing the unaltered imaging data as acquired during clinical examination. The second column displays the preprocessed images, in which intensity normalization, artifact suppression, and contrast enhancement are applied to improve visibility of anatomical structures and pathological regions. In the third column, the manually annotated ground truth (GT) masks are shown in green, representing expert-defined lesion boundaries verified through radiological consensus. The fourth column depicts the segmentation results produced by the proposed PM1 model, where predicted lesion regions are highlighted in red. Finally, the fifth column overlays the GT masks onto the PM1 predictions, enabling direct visual assessment of prediction accuracy. This overlay highlights both the areas of precise alignment and the locations of minor deviations, offering valuable insights for targeted model refinement. The visual progression in Fig. 8 demonstrates the PM1 model's capacity to accurately approximate expert annotations, while also emphasizing its robustness in handling variations in lesion size, shape, and contrast. This comprehensive depiction supports the quantitative findings and confirms the method's clinical applicability for reliable stroke lesion localization.

TABLE I. QUANTITATIVE COMPARISON OF THE PROPOSED FEATURE PYRAMID NETWORK WITH DUAL-DECODER SUPERVISION AGAINST BASELINE AND STATE-OF-THE-ART METHODS FOR STROKE LESION LOCALIZATION IN MULTI-MODAL BRAIN MRI

Method	DSC	Jaccard	Precision	Recall	HD95	AUC
Proposed: FPN + Dual-Decoder + Proposed: FPN + Dual-Decoder + Aux Supervision	**0.873**	**0.754**	**0.879**	**0.861**	**6.54**	**0.918**
U-Net (baseline)	0.794	0.663	0.801	0.782	8.41	0.872
Attention U-Net	0.816	0.687	0.826	0.803	7.95	0.884
Swin-UNet	0.832	0.705	0.843	0.817	7.51	0.893
FPN + Single Decoder	0.846	0.721	0.854	0.829	7.18	0.902

Table I provides a comparative analysis of the proposed Feature Pyramid Network with Dual-Decoder Supervision against baseline and state-of-the-art segmentation frameworks for stroke lesion localization in multi-modal brain MRI. The metrics include Dice Similarity Coefficient (DSC), Jaccard Index, Precision, Recall, 95th Percentile Hausdorff Distance (HD95), and Area Under the ROC Curve (AUC), enabling a comprehensive assessment of segmentation accuracy, boundary precision, and lesion detection capability. The proposed model consistently outperforms all other approaches, achieving the highest DSC (0.873) and Jaccard Index (0.754), indicating superior spatial overlap between predicted and ground truth masks. Precision and Recall values of 0.879 and 0.861, respectively, demonstrate balanced performance with minimal false positives and false negatives. The lowest HD95 (6.54) reflects improved boundary alignment, while an AUC of 0.918 confirms robust lesion presence classification. These results highlight the synergistic benefits of multi-scale feature fusion, dual-decoder specialization, and auxiliary supervision in enhancing both global context understanding and local detail preservation. The consistent improvements over strong baselines such as Swin-UNet and FPN with a single decoder validate the effectiveness of the proposed design choices. Overall, Table I underscores the framework's capacity to deliver accurate, reliable, and clinically meaningful segmentation performance.

V. DISCUSSION

A. Performance Evaluation

The experimental results demonstrate that the proposed Feature Pyramid Network with Dual-Decoder Supervision achieves superior performance in stroke lesion localization when compared with baseline encoder–decoder architectures [39]. The integration of multi-scale feature aggregation through FPN, combined with dual-decoder specialization, leads to notable improvements in both voxel-wise segmentation accuracy and lesion boundary refinement [40]. Quantitative

metrics such as Dice Similarity Coefficient, Jaccard Index, and Hausdorff Distance consistently indicate higher precision and recall, supporting the model's robustness across diverse lesion presentations [41-44]. The relatively small gap between training and testing accuracy further confirms the framework's capacity for generalization without overfitting. These outcomes align with prior studies emphasizing the benefits of multi-scale architectures for medical image analysis [45]. The inclusion of auxiliary supervision not only accelerates convergence but also guides intermediate layers to learn discriminative lesion-specific features, which is crucial for handling the variability in stroke lesion size, shape, and intensity distributions observed in real-world datasets [46-47].

B. Impact of Multi-Modal Integration

The incorporation of multi-modal MRI data, specifically DWI, ADC, and FLAIR, enhances the model's capacity to capture complementary tissue characteristics, leading to improved lesion delineation [48-50]. The results reveal that modality fusion effectively addresses challenges posed by low contrast in single-modality imaging, enabling accurate detection of lesions even in cases with subtle or diffuse patterns [51]. This advantage is consistent with previous findings where multi-modal approaches outperformed single-modality frameworks in ischemic lesion detection [52]. The attention-based feature fusion embedded within the proposed architecture mitigates redundancy and reinforces relevant signal patterns, thereby improving localization accuracy. Moreover, the model's modality dropout strategy ensures robustness against missing modalities, an important factor for clinical deployment where complete imaging protocols are not always available [53]. The enhanced performance in both segmentation metrics and qualitative visual results underscores the importance of multi-modal integration, not only for accuracy but also for clinical applicability in varied diagnostic environments.

C. Effectiveness of Dual-Decoder Supervision

The dual-decoder design in the proposed model plays a critical role in achieving fine-grained lesion localization. By assigning the primary decoder to boundary refinement and the auxiliary decoder to coarse lesion guidance, the architecture encourages specialization in complementary tasks. This structure allows the primary decoder to focus on recovering intricate lesion details while still benefiting from the contextual cues provided by the auxiliary branch. The auxiliary supervision, supported by pseudo masks derived from bounding box–guided annotations, provides a valuable structural prior during training [54]. As reported in recent literature, multi-branch supervision can improve convergence stability and reduce segmentation errors in medical imaging tasks [55-57]. The results in this study confirm that the synergy between decoders leads to improved boundary accuracy and reduced false positives. Furthermore, the qualitative results demonstrate that even when pseudo masks contain inaccuracies, the network refines these into precise voxel-level predictions, reinforcing the utility of this multi-task learning paradigm.

D. Model Generalization and Robustness

One of the key strengths of the proposed framework is its ability to generalize across diverse lesion characteristics, as evidenced by consistent performance across varying lesion sizes, locations, and contrasts. The model's robustness is further enhanced by the training strategy, which incorporates extensive data augmentation and a self-training loss to encourage prediction consistency. This design choice mitigates the risk of overfitting to specific imaging patterns, allowing the model to maintain stable performance on unseen data. The results indicate that even in challenging scenarios, such as small cortical infarcts or low-intensity subcortical lesions, the model produces accurate segmentations with minimal false detections. Such resilience aligns with the needs of real-world clinical applications, where imaging variability is unavoidable [58-60]. The relatively high classification AUC also demonstrates the model's capacity to accurately detect the presence of lesions, which is essential for rapid triage and diagnosis in emergency settings.

E. Clinical Implications and Future Work

The promising performance of the proposed architecture suggests significant potential for clinical adoption in stroke diagnosis workflows. Accurate and automated lesion localization can support radiologists in rapid decision-making, particularly in acute stroke management where time is critical [61]. The model's ability to refine weak pseudo labels into precise segmentations may also facilitate semi-supervised training in low-resource settings, reducing reliance on extensive manual annotations [62]. Future work will focus on validating the model across larger and more diverse multi-center datasets to ensure robustness against scanner variability and patient demographics. Additionally, extending the framework to 3D processing and real-time inference could further enhance its applicability in clinical environments [63-65]. The integration of explainable AI techniques [66] would also improve clinician trust by providing transparent visualizations of model decisions, an increasingly important factor for regulatory approval and ethical deployment.

VI. Conclusion

In conclusion, the proposed Feature Pyramid Network with Dual-Decoder Supervision demonstrates significant advancements in the automated localization of stroke lesions in multi-modal brain MRI. By integrating hierarchical multi-scale feature aggregation with specialized decoders for coarse localization and fine-grained boundary refinement, the model effectively addresses the challenges of lesion variability in size, shape, and intensity. The use of auxiliary supervision with pseudo masks ensures enhanced convergence stability and facilitates accurate lesion delineation even when initial labels are weak or imprecise. Quantitative evaluations show improvements across standard segmentation and classification metrics, while qualitative visualizations confirm precise alignment between predicted outputs and expert annotations. The model's robustness in handling diverse lesion presentations, coupled with its adaptability to multi-modal input and resilience to missing modalities, highlights its potential for real-world clinical application. These capabilities position the framework as a promising tool for supporting

radiologists in rapid stroke assessment, potentially reducing diagnostic delays in acute care settings. Future work should focus on expanding validation to multi-center datasets, integrating explainable AI techniques for clinical trust, and exploring 3D and real-time processing to further enhance diagnostic accuracy and workflow efficiency. This research lays a strong foundation for scalable, accurate, and interpretable AI-driven stroke imaging solutions.

ACKNOWLEDGMENTS

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”.

REFERENCES

[1] Yu, H., Wang, Z., Sun, Y., Bo, W., Duan, K., Song, C., ... & Wu, N. (2023). Prognosis of ischemic stroke predicted by machine learning based on multi-modal MRI radiomics. Frontiers in Psychiatry, 13, 1105496.

[2] Alkaraawi, M. R. L., & Shengwu, X. (2025). MMG-SiamNet: multi-modal granular siamese network for robust ischemic stroke detection via cross-modal learning. Signal, Image and Video Processing, 19(11), 934.

[3] Yoon, C., Misra, S., Kim, K. J., Kim, C., & Kim, B. J. (2023). Collaborative multi-modal deep learning and radiomic features for classification of strokes within 6 h. Expert Liu, S., Zhang, B., Fang, R., Rueckert, D., & Zimmer, V. A. (2023, October). Dynamic graph neural representation based multi-modal fusion model for cognitive outcome prediction in stroke cases. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 338-347). Cham: Springer Nature Switzerland.Systems with Applications, 228, 120473.

[4] Zhao, X., Li, S., Jiang, L., & Xu, M. (2024, November). A New Multi-modal Dataset and Two-stage DNN Approach for Acute Ischemic Stroke Detection. In 2024 IEEE International Conference on Signal, Information and Data Processing (ICSIDP) (pp. 1-5). IEEE.

[5] Omarov, Batyrkhan; Suliman, Azizah; Kushibar, Kaisar. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology; Islamabad 91.2 (Sep 2016): 238-248

[6] Liao, W., Jiang, P., Lv, Y., Xue, Y., Chen, Z., & Li, X. (2023, April). MCRLe: Multi-Modal Contrastive Representation Learning For Stroke Onset Time Diagnosis. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.

[7] Delgrange, C., Demler, O., Mora, S., Menze, B., de la Rosa, E., & Davoudi, N. (2024). A Self-Supervised Model for Multi-modal Stroke Risk Prediction. arXiv preprint arXiv:2411.09822.

[8] Blessa Binolin Pepsi M., Anandhi H., Karunyaharini S., Visali N., "Convolutional Neural Network-based Stacking Technique for Brain Tumor Classification using Red Panda Optimization", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.52-67, 2025. DOI:10.5815/ijitcs.2025.05.05

[9] Tsai, C. L., Su, H. Y., Sung, S. F., Lin, W. Y., Su, Y. Y., Yang, T. H., & Mai, M. L. (2024). Fusion of diffusion weighted MRI and clinical data for predicting functional outcome after acute ischemic stroke with deep contrastive learning. arXiv preprint arXiv:2402.10894.

[10] Chaithra, I. V., Bhavana, L. U., Bhavani, K. G., Sinchana, E., Rakshitha, K. A., & Mithuna, B. N. (2025, April). Stroke Detection and Prediction Using Deep Learning Techniques. In 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS) (pp. 1-6). IEEE.

[11] Chen, S., Zhao, X., Duan, Y., Ju, R., Zang, P., & Qi, S. (2025). M2FNet: multi-modality multi-level fusion network for segmentation of acute and sub-acute ischemic stroke. Complex & Intelligent Systems, 11(6), 1-24.

[12] Zhang, Z., Ding, Z., Chen, F., Hua, R., Wu, J., Shen, Z., ... & Xu, X. (2024). Quantitative analysis of multimodal MRI markers and clinical risk factors for cerebral small vessel disease based on deep learning. International Journal of General Medicine, 739-750.

[13] Tuxunjiang, P., Huang, C., Zhou, Z., Zhao, W., Han, B., Tan, W., ... & Wang, Y. (2025). Prediction of NIHSS Scores and Acute Ischemic Stroke Severity Using a Cross-attention Vision Transformer Model with Multimodal MRI. Academic Radiology.

[14] Sun, M., Li, X., & Sun, W. (2024). Image generation and lesion segmentation of brain tumors and stroke based on GAN and 3D ResU-Net. IEEE Access.

[15] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., Ray, S., ... & Wang, C. (2023). A computer vision-based lane detection technique using gradient threshold and hue-lightness-saturation value for an autonomous vehicle. International Journal of Electrical and Computer Engineering, 13(1), 347.

[16] Salvalaggio, S., Turolla, A., Andò, M., Barresi, R., Burgio, F., Busan, P., ... & Filippini, N. (2023). Prediction of rehabilitation induced motor recovery after stroke using a multi-dimensional and multi-modal approach. Frontiers in Aging Neuroscience, 15, 1205063.

[17] Fleury, L., Koch, P. J., Wessel, M. J., Bonvin, C., San Millan, D., Constantin, C., ... & Hummel, F. C. (2022). Toward individualized medicine in stroke—The TiMeS project: Protocol of longitudinal, multi-modal,multi-domain study in stroke. Frontiers in neurology, 13, 939640.

[18] Pani, K., & Chawla, I. (2025). Fuzzy Inception UNet: Bridging Uncertainties in MRI for Multi-Modal Brain Tumor Segmentation. SN Computer Science, 6(6), 621.

[19] Xin, N. I. E. (2025). Multi-Modal Image Fusion for Medical Diagnosis: Combining MRI And CT Using Deep Generative Models. Clinical Medicine And Health Research Journal, 5(03), 1313-1327.

[20] von Braun, M. S., Starke, K., Peter, L., Kürsten, D., Welle, F., Schneider, H. R., ... & Saur, D. (2025). Prediction of tissue and clinical thrombectomy outcome in acute ischaemic stroke using deep learning. Brain, awaf013.

[21] Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., & Wang, X. (2025). Multi-modal Vision Pre-training for Medical Image Analysis. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 5164-5174).

[22] Aschalew Arega, Durga Prasad Sharma, "Evaluating Energy-efficiency and Performance of Cloud-based Healthcare Systems Using Power-aware Algorithms: An Experimental Simulation Approach for Public Hospitals", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.3, pp.72-96, 2025. DOI:10.5815/ijitcs.2025.03.06

[23] Omarov, B., Altayeva, A., & Cho, Y. I. (2017, May). Smart building climate control considering indoor and outdoor parameters. In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 412-422). Cham: Springer International Publishing.

[24] Aouthu, S., Suman, S. K., Anuradha, S., Sanapala, R. K., & Geetha, A. (2025). Automated Diagnosis of Acute Cerebral Ischemic Stroke Lesions using Capsule Graph Neural Networks from Diffusion-weighted MRI Scans. Journal of Electrical Engineering & Technology, 20(4), 2631-2650.

[25] Abujaber, A. A., Albalkhi, I., Imam, Y., Yaseen, S., Nashwan, A. J., Akhtar, N., & Alkhawaldeh, I. M. (2025). Machine learning-based prediction of 90-day prognosis and in-hospital mortality in hemorrhagic stroke patients. Scientific Reports, 15(1), 16242.

[26] Gulli, G., Colman, J., Duan, W., Ye, X., & Zhang, L. Multi-Modal Brain Segmentation Using Hyper-Fused Convolutional Neural Network.

[27] Duan, W., Zhang, L., Colman, J., Gulli, G., & Ye, X. (2022, September). MidFusNet: Mid-dense Fusion Network for Multi-modal Brain MRI Segmentation. In International MICCAI Brainlesion Workshop (pp. 102-114). Cham: Springer Nature Switzerland.

[28] Altayeva, A. B., Omarov, B. S., Aitmagambetov, A. Z., Kendzhaeva, B. B., & Burkitbayeva, M. A. (2014). Modeling and exploring base station characteristics of LTE mobile networks. Life Science Journal, 11(6), 227-233.

[29] Liu, Y., Tang, Y., Li, Z., Yu, P., Yuan, J., Zeng, L., ... & Zhao, L. (2025). Prediction of clinical efficacy of acupuncture intervention on upper limb dysfunction after ischemic stroke based on machine learning:

a study driven by DSA diagnostic reports data. Frontiers in Neurology, 15, 1441886.

[30] Kumar, P., Mohammad, T. K., Kumar, A. P., Kassym, R., AS, T., & Akmaral, T. (2024, November). Synthesizing Multi-Modal Imaging for Enhanced Brain Mapping in Neurology: A State-of-the-art Review. In 2024 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES) (pp. 1-6). IEEE.

[31] Seth, P., & Khan, A. (2024, December). ReFuSeg: Regularized Multi-modal Fusion for Precise Brain Tumour Segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 9th International Workshop, BrainLes 2023, and 3rd International Workshop, SWITCH 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8 and 12, 2023, Revised Selected Papers (Vol. 14668, p. 69). Springer Nature.

[32] Yue, G., Zhuo, G., Zhou, T., Liu, W., Wang, T., & Jiang, Q. (2023). Adaptive cross-feature fusion network with inconsistency guidance for multi-modal brain tumor segmentation. IEEE Journal of Biomedical and Health Informatics.

[33] Hossein. Abbasi, Ahmed. Alshaeeb, Yasin. Orouskhani, Behrouz. Rahimi, Mostafa. Shomalzadeh, "Advanced Deep Learning Models for Accurate Retinal Disease State Detection", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.3, pp.61-71, 2024. DOI:10.5815/ijitcs.2024.03.06

[34] Chen, S., Zhao, S., & Lan, Q. (2022). Residual block based nested U-type architecture for multi-modal brain tumor image segmentation. Frontiers in Neuroscience, 16, 832824.

[35] Yang, H., Zhou, T., Zhou, Y., Zhang, Y., & Fu, H. (2023). Flexible fusion network for multi-modal brain tumor segmentation. IEEE journal of biomedical and health informatics, 27(7), 3349-3359.

[36] Wang, Z., Zhu, H., Huang, B., Wang, Z., Lu, W., Chen, N., & Wang, Y. (2023). M-MSSEU: source-free domain adaptation for multi-modal stroke lesion segmentation using shadowed sets and evidential uncertainty. Health Information Science and Systems, 11(1), 46.

[37] Abdmouleh, N., Echtioui, A., Kallel, F., & Hamida, A. B. (2022, December). Modified u-net architeture based ischemic stroke lesions segmentation. In 2022 IEEE 21st international Ccnference on Sciences and Techniques of Automatic Control and Computer Engineering (STA) (pp. 361-365). IEEE.

[38] Zubair, M., Hussai, M., Al-Bashrawi, M. A., Bendechache, M., & Owais, M. (2025). A Comprehensive Review of Techniques, Algorithms, Advancements, Challenges, and Clinical Applications of Multi-modal Medical Image Fusion for Improved Diagnosis. arXiv preprint arXiv:2505.14715.

[39] Xiang, H., Xiong, Y., Shen, Y., Li, J., & Liu, D. (2025). A collaborative multi-task model for immunohistochemical molecular sub-types of multi-modal breast cancer MRI images. Biomedical Signal Processing and Control, 100, 107137.

[40] Shi, X., Jain, R. K., Li, Y., Chai, S., Cheng, J., Bai, J., ... & Chen, Y. W. (2025). Multi-modal Medical SAM: An Adaptation Method of Segment Anything Model (SAM) for Glioma Segmentation Using Multi-modal MR Images. ACM Transactions on Computing for Healthcare, 6(2), 1-21.

[41] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[42] An Cong Tran, Huynh Vo-Thuy, "A ViT-based Model for Detecting Kidney Stones in Coronal CT Images", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.1-11, 2025. DOI:10.5815/ijitcs.2025.05.01

[43] Fang, E., Fartaria, M. J., Ann, C. N., Maréchal, B., Kober, T., Lim, J. X., ... & Chan, L. L. (2021). Clinical correlates of white matter lesions in Parkinson's disease using automated multi-modal segmentation measures. Journal of the neurological sciences, 427, 117518.

[44] Kasliwal, A., Sagaram, S., Srivastava, L., Seth, P., & Khan, A. (2023, October). ReFuSeg: Regularized Multi-modal Fusion for Precise Brain Tumour Segmentation. In International MICCAI Brainlesion Workshop (pp. 69-80). Cham: Springer Nature Switzerland.

[45] Karimzadeh, M., Seyedarabi, H., Jodeiri, A., & Afrouzian, R. (2025). Enhanced Brain Stroke Lesion Segmentation in MRI Using a 2.5 D Transformer Backbone U-Net Model. Brain Sciences, 15(8), 778.

[46] Zhang, G., Gao, Z., Duan, C., Liu, J., Lizhu, Y., Liu, Y., ... & Dai, Q. (2025). A Multimodal Vision-text AI Copilot for Brain Disease Diagnosis and Medical Imaging. medRxiv, 2025-01.

[47] Akbari, B., Huber, B. R., & Sherman, J. H. (2025). Unlocking the hidden depths: multi-modal integration of imaging mass spectrometry-based and molecular imaging techniques. Critical Reviews in Analytical Chemistry, 55(1), 109-138.

[48] Sharma, V., Vilarrubias, R. B., & Verschure, P. F. (2023, July). BrainX3: a neuroinformatic tool for interactive exploration of multimodal brain datasets. In Conference on Biomimetic and Biohybrid Systems (pp. 157-177). Cham: Springer Nature Switzerland.

[49] Jin, W., Li, X., & Hamarneh, G. (2022, June). Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 11945-11953).

[50] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble Approach Combining Deep Residual Networks and BiGRU with Attention Mechanism for Classification of Heart Arrhythmias. Computers, Materials & Continua, 80(1).

[51] Cai, T., Wong, K., Wang, J. Z., Huang, S., Yu, X., Volpi, J. J., & Wong, S. T. (2024, November). M 3 Stroke: Multi-Modal Mobile AI for Emergency Triage of Mild to Moderate Acute Strokes. In 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1-8). IEEE.

[52] Singh, R., & Lee, E. Y. P. (2024, November). Multi-Modal MRI Brain Tumor Segmentation Using Attention Modules. In 2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI) (pp. 338-342). IEEE.

[53] Shourove Sutradhar Dip, Md. Habibur Rahman, Nazrul Islam, Md. Easin Arafat, Pulak Kanti Bhowmick, Mohammad Abu Yousuf, "Enhancing Brain Tumor Classification in MRI: Leveraging Deep Convolutional Neural Networks for Improved Accuracy", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.3, pp.12-21, 2024. DOI:10.5815/ijitcs.2024.03.02

[54] Altayeva, A., Omarov, B., Jeong, H. C., & Im Cho, Y. (2016). Multi-step face recognition for improving face detection and recognition rate. Far East Journal of Electronics and Communications, 16(3), 471.

[55] Luo, Y., Li, C., Sun, Y., & Fan, H. (2022, November). Multi-Modal Magnetic Resonance Images Segmentation Based on An Improved 3DUNet. In 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-6). IEEE.

[56] Su, F., Yi, X., Cheng, Y., Ma, Y., Zu, W., Zhao, Q., ... & Ma, L. (2025). From Slices to Volumes: A Scalable Pipeline for Developing General-Purpose Brain MRI Foundation Models. medRxiv, 2025-04.

[57] Rahman, A., Chowdhury, M. E., Wadud, M. S. I., Sarmun, R., Mushtak, A., Zoghoul, S. B., & Al-Hashimi, I. (2025). Deep learning-driven segmentation of ischemic stroke lesions using multi-channel MRI. Biomedical Signal Processing and Control, 105, 107676.

[58] Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., ... & Pei, Y. (2024). A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. Bioengineering, 11(3), 219.

[59] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[60] Kumar, A. (2022). Deep learning for multi-modal medical imaging fusion: Enhancing diagnostic accuracy in complex disease detection. Int J Eng Technol Res Manag, 6(11), 183.

[61] Sinha, S., Bhatt, M., Anand, A., & Areeckal, A. S. (2024). EnigmaNet: A Novel Attention-Enhanced Segmentation Framework for Ischemic Stroke Lesion Detection in Brain MRI. IEEE Access, 12, 91480-91498.

[62] Chen, J., Huang, G., Yuan, X., Zhong, G., Zheng, Z., Pun, C. M., ... & Huang, Z. (2023). Quaternion cross-modality spatial learning for multi-

modal medical image segmentation. IEEE Journal of Biomedical and Health Informatics, 28(3), 1412-1423.

[63] Mathew, P. S., Pillai, A. S., di Biase, L., & Abraham, A. (2025). Swin-BSSeg: A Novel Swin Transformer-Enhanced Architecture for Accurate Ischemic Stroke Lesion Segmentation in MRI Images. International Journal of Online & Biomedical Engineering, 21(9).

[64] Inamdar, M. A., Gudigar, A., Raghavendra, U., Salvi, M., Aman, R. R. A. B. R., Gowdh, N. F. M., ... & Acharya, U. R. (2025). A Dual-Stream Deep Learning Architecture With Adaptive Random Vector Functional Link for Multi-Center Ischemic Stroke Classification. IEEE Access.

[65] Tang, T., Cui, Y., Lu, C., Li, H., Zhou, J., Zhang, X., ... & Ju, S. (2025). Evaluating Performance of a Deep Learning Multilabel Segmentation Model to Quantify Acute and Chronic Brain Lesions at MRI after Stroke and Predict Prognosis. Radiology: Artificial Intelligence, 7(3), e240072.

[66] Ince, S., Kunduracioglu, I., Algarni, A., Bayram, B., & Pacal, I. (2025). Deep learning for cerebral vascular occlusion segmentation: a novel ConvNeXtV2 and GRN-integrated U-Net framework for diffusion-weighted imaging. Neuroscience, 574, 42-53.