Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Benchmarking Medical Segmentation Architectures: A Systematic Review and Comparative Analysis of CNN, Transformer, MLP, and Foundation Models

Shreyansh Palwalia[1], Umesh Patil[1], Dr. Sonika Dahiya[2]

[1]Student, Department of Software Engineering, Delhi Technological University
[2]Assistant Professor, Department of Software Engineering, Delhi Technological University

Abstract - Medical image segmentation has experienced a revolutionary evolution from traditional Convolutional Neu-ral Networks (CNNs) to modern methodologies such as Transformers, Multi-Layer Perceptrons (MLPs), and Foundation Models. Although quantitative performance metrics continue to improve, the comparative litera-ture often neglects practical deployment constraints, specifically inference latency and computational cost. This investigation combines a Systematic Literature Review (SLR) synthesizing 28 peer-reviewed articles (2021–2025) with rigorous experimental benchmarking. Eight representative models, including RepLKNet, Swin-UNETR, AS-MLP, and WS-ICL, were evaluated on a standardized NVIDIA RTX 3050 workstation to assess real-world efficiency. The benchmarking identified a "Lightweight Paradox," wherein theoretically compact MLP models (e.g., UNeXt) demonstrated higher latency than conventional CNNs due to suboptimal hardware implementa-tions of token-shifting operations. Furthermore, while Foundation Models exhibit superior generalization, their prohibitive computational requirements (exceeding 1,000 GFLOPs) pose severe challenges for real-time clinical application. We conclude that while Transformers continue to achieve state-of-the-art 3D volumetric accu-racy, optimized MLP-based architectures and standard CNNs currently offer the most viable efficiency-accuracy trade-offs for mobile health and resource-constrained deployments.

Keywords: Medical Image Segmentation, Deep Learning, Benchmarking, Foundation Models, Efficiency Anal-ysis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

1 Introduction

Medical imaging techniques such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound are essential elements of a modern diagnostic regimen [1, 32]. However, the manual interpretation of these images is still a difficult an d hi ghly observer-dependent task [2]. Therefore, automated medical image segmentation, which refers to the classification of individual pixels to delimit anatomical structures, has become a critical component of computer-aided diagnostic systems [31].

The recent architectural evolution in segmentation model development has progressed rapidly, transitioning from foundational CNNs to advanced hybrid and Transformer-based designs [33]. CNNs and in particular U-Net have formed a formidable basis of feature extraction [3]. Nevertheless, the inference capability of CNNs usually shows a relatively slow speed in detecting long-range dependencies, and as an alternative approach, Vision Transformer (ViT), e.g. SwinUNETR, use self-attention mechanisms to incorporate global context [10]. At the same time, multilayer perceptron (MLP) based models such as UNeXt and AS-MLP have been created, which have lighter computational footprints [18]. Recently, foundation models (e.g., WS-ICL, SAM) have put more emphasis on prompt-based zero-shot generalization [23].

In spite of these developments, there remains a critical gap in the literature: the lack of hardware-consistent benchmarking. The current reviews largely focus on the accuracy measures, e.g., the Dice coefficient, but often do not consider practical implementation issues, such as the inference latency, floating-point operations (FLOPs) and memory requirements [12].In a clinical scenario where resources are limited, e.g. a district hospital or a moving unit providing diagnostic system, a model giving only marginal gains in performance may be shown ineffectual w hen it r equires server-grade GPUs. Moreover, the cross-study comparisons of the models are usually compromised by the differences in the hardware specifications and preprocessing pipelines.

The current research will be a fusion of a systematic literature review (SLR) of 28 modern research (2021-2025) and a controlled experimental benchmarking model to overcome these obstacles. We compare the representative models of four families, which are CNNs, Transformers, MLPs, and foundation models, and one common hardware and configuration (NVIDIA RTX 3050). Our most important contribution is the fully developed accuracy -versus- efficiency analysis which provides evidence based advice on choosing architectures appropriate to particular deployment situations.

2 Methodology

2.1 Systematic Literature Review (SLR)

To ensure a structured assessment of architectural advancements, this study followed the PRISMA 2020 guidelines (see Figure 1). A complete search was undertaken in IEEE Xplore, PubMed, Scopus, and arXiv of studies published in January 2021 to December 2025. Search queries included terms that are relevant to segmentation procedures and model families such as "medical image segmentation," Transformer, MLP, and Foundation Model.

The initial search yielded 255 records. After removing duplicates and applying exclusion criteria, such as non-medical domains and studies lacking quantitative measurements, 28 papers were chosen for final inclusion. We systematically reviewed 28 papers, but our entire manuscript cites 40 papers to provide proper background and context. These selected studies encompass a broad spectrum of state-of-the-art segmentation architectures, which we categorized into four distinct families: advanced Convolutional Neural Networks [5, 6, 7, 8, 9, 20], Vision Transformers (a focus justified by recent comprehensive surveys [40]) [13, 14, 15, 16], MLP-based frameworks [19, 21], and recent Foundation Models [24, 25, 26, 27, 28, 29].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

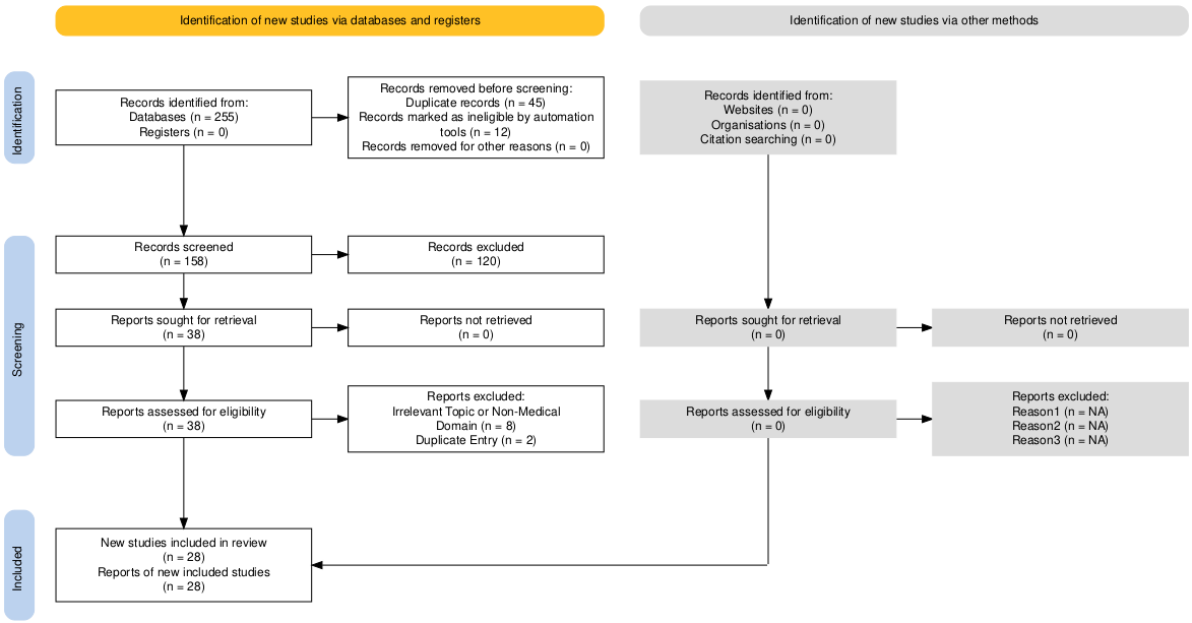

Figure 1: PRISMA 2020 flow diagram that outlines the selection of the literature.

2.2 Review of Architectural Families

Based on the 28 selected studies, the deep learning architectures for medical image segmentation were categorized into four distinct families: Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Multi-Layer Perceptrons (MLPs), and Foundation Models. This taxonomy is illustrated in Figure 2, and a brief summary of the eight representative models used in the benchmarking process is provided in Table 1.

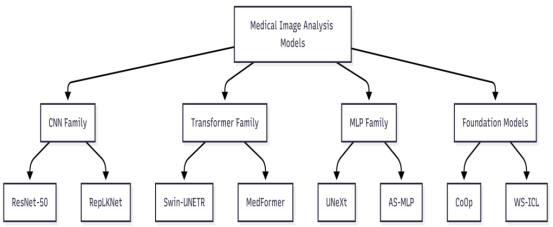

Figure 2: Taxonomy of deep learning architectures that were evaluated in the framework of this study.

2.2.1 Convolutional Neural Networks (CNNs)

CNNs have traditionally dominated medical segmentation due to their efficiency in local spatial pattern extraction [8, 9]. Although the encoder-decoder architecture of 2D U-Net is a natural baseline, extensions like 3D U-Net have been pivotal for learning dense volumetric segmentation from sparse annotations [35, 34, 20]. Furthermore, new architectures including RepLKNet have been used to scale the size of convolutional kernels to 31×31 to replicate the effects of the global receptive field of Transformers at the same hardware cost [4].

2.2.2 Vision Transformers (ViTs)

Vision Transformers, originally introduced for large-scale image recognition [36], use self-attention to overcome the shortcoming of CNNs in representations of long-range dependencies [10, 11]. Such architectures as Swin-UNETR are adapting hierarchical Transformers on 3D volumetric segmentation, exhibiting a better global understanding of the context, especially when enhanced by self-supervised pre-training [10, 38]. However, such an improvement usually involves a quadratic cost in terms of computation [10, 12].

2.2.3 Multi-Layer Perceptrons (MLPs)

Transformer computational complexity has caused the introduction of MLP based architectures, which replace attention layers with token mixing operations of simplified nature [12, 13]. Models like UNeXt and AS-MLP are designed to be extremely light-weight in terms of the number of parameters and therefore potentially with accelerated inference on resource-constrained devices [13, 14].

2.2.4 Foundation Models

The current paradigm shift being explored is that of large, pre-trained models, specifically designed to facilitate zero-shot or weakly-supervised generalization across an array of domains [15, 16]. Extensive experimental studies reveal that while models like SAM show impressive zero-shot capabilities, specialized adaptations such as the Medical SAM Adapter are required to efficiently bridge the gap for complex medical modalities [25, 37].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Table 1: Summary of representative models of the chosen and their salient characteristics.

Reference	Year	Family	Model	Key Contribution
Ding et al. [4]	2022	CNN	RepLKNet	Revisiting the work on large-kernel designs with size (31×31) to use as a competitive receptive field to Transformer architectures.
He et al. [30]	2016	CNN	ResNet-50	The standard residual network baseline is commonly used as an encoder backbone.
Hatamizadeh et al. [10]	2022	Transformer	Swin-UNETR	The hierarchical Swin Transformer encoder is modified to application in three dimensional volumetric segmentation.
Xia et al. [11]	2023	Transformer	MedFormer	Transformer multi-grained feature integration on medical data.
Valanarasu et al. [18]	2022	MLP	UNeXt	The use of a convolution-MLP hybrid that uses tokenized MLPs is explored to achieve ultra-fast medical segmentation.
Lian et al. [17]	2021	MLP	AS-MLP	Axial- Shifted MLP that allows the existence of spatial inter-nactions in the absence of attention mechanisms.
Zhou et al. [22]	2022	Foundation	CoOp	Context Optimization for adapting vision-language models to downstream tasks.
Hu et al. [23]	2024	Foundation	WS-ICL	Weakly-Supervised In-Context Learning for segmentation using sparse annotations.

2.3 Experimental Benchmarking Framework

Unlike prior works that rely on reported metrics from diverse hardware environments, this study evaluates all models on a unified testbed to ensure fair comparison.

2.3.1 Hardware Environment

All experiments were executed on a consumer-grade workstation to simulate resource-constrained clinical settings. The specifications included:

- **GPU:** NVIDIA GeForce RTX 3050 Laptop GPU (4 GB VRAM)
- **CPU:** AMD Ryzen 7 5800H (8 Cores)
- **Framework:** PyTorch 2.1 with CUDA 12.1 and MONAI 1.3

2.3.2 Dataset Selection

To evaluate generalization across modalities, three diverse datasets were selected:

1. **ISIC 2018 (Dermoscopy):** 2,594 images resized to 256×256 for 2D lesion segmentation.
2. **BUSI (Ultrasound):** 780 images resized to 224×224 for tumor classification.
3. **BraTS 2021 (MRI):** Multi-modal 3D volumetric data, processed as $96 \times 96 \times 96$ patches.

3 Results

3.1 Quantitative Benchmarking

Table 2 and Figure 3 present the consolidated benchmarking results for all eight models evaluated on the NVIDIA RTX 3050 GPU.

The results indicate great trade-offs between architectural families:

- **Efficiency:** AS- MLP has high efficiency, with the lowest FLOPs (0.22 M) and parameter number (0.08 M), making it very suitable for mobile platforms.
- **The Lightweight Paradox:** Once more, as expected by theory, UNeXt incurs the longest inference latency of 2D models although with the fewest number of parameters (7.76M), with a latency of 628ms. This observation, however, indicates that, the token-shifting operations with no substantial optimizations, place an enormous load on the memory-access cost on the traditional GPUs.
- **Context Cost:** Swin-UNETR has powerful 3D modelling features but it requires 85.5 Gflops per patch, indicating that the computational cost of self attention mechanisms for volumetric tasks is very high.
- **Foundation Overhead:** The collective overhead for the foundation compute in WS-ICL is the highest (1092 GFLOPs). This confirms that in-context learning shifts the memory use burden of train time to inference time.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Table 2: Comparative analysis of speed of inference and computational complexity and size of models.

Family	Model	Input	Time (ms) ↓	GFLOPs ↓	Params (M) ↓
CNN	RepLKNet [4]	2D	346.27	15.60	79.86
CNN	ResNet-50 [30]	2D	21.82	4.13	25.56
Transformer	Swin-UNETR [10]	3D	214.57	85.51	15.51
Transformer	MedFormer [11]	2D	114.55	2.95	17.31
MLP	UNeXt [18]	2D	628.64	18.08	7.76
MLP	AS-MLP [17]	2D	**190.71**	**0.22**	**0.08**
Foundation	CoOp [22]	2D	22.21	4.13	25.56
Foundation	WS-ICL [23]	3D	472.05	1092.44	0.86

Note: Time averages were found over 10 synchronous runs; the most efficient metrics are shown in bold type.

3.2 Accuracy Metrics (Literature Baselines)

To properly contextualize the computational efficiency results, it is imperative to evaluate the corresponding diagnostic accuracy. Because foundational training protocols and dataset splits differ among the original papers, Table 3 presents the baseline level of accuracy (Dice Similarity Coefficient or Accuracy) as reported in the literature.

Table 3: Reported accuracy measures from baselines of original literature.

Family	Model	Dataset (Task)	Reported Metric
CNN	RepLKNet	Cityscapes/Medical (Seg)	Dice: 0.890
CNN	ResNet-50	ISIC 2018 (Class)	Acc: 0.935
Transformer	Swin-UNETR	BraTS 2021 (3D Seg)	Dice: 0.881
Transformer	MedFormer	ISIC 2018 (Seg)	Dice: 0.871
MLP	AS-MLP	ISIC 2018 (Seg)	Dice: 0.862
MLP	UNeXt	ISIC 2018 (Seg)	Dice: 0.855
Foundation	CoOp	11-Datasets (Class)	Acc: 0.793
Foundation	WS-ICL	Abdominal MRI (Seg)	Dice: 0.825

These outcomes are reported metrics that prove that although Transformer models like SwinUNETR can have state-of-the-art volumetric precision (Dice: 0.881) [10], MLP based models like AS-MLP [17] are still very competitive (Dice: 0.862) on 2D tasks despite their minimal computational footprint. This observation directly shows the trade-off that is required in the choice of models when deploying limited resources.

3.3 Qualitative Analysis

In line with qualitative findings (Figure 4), lightweight models like UNeXt are still able to create sharp boundaries of segmentations even though they have a simplified architecture.

4 Discussion

The benchmarking results highlight a complex landscape where architectural efficiency does not always align with parameter counts. This section interprets the "Efficiency-Accuracy" trade-offs observed.

4.1 The Lightweight Paradox

An important observation of this study is that model size-latency of inference does not match MLP based architectures. While UNeXt is the smallest model evaluated (7.76 M parameters), it exhibited significantly higher latency (628 ms) compared to the much larger RepLKNet (346 ms) and ResNet-50 (21 ms). This finding is analogous to the so-called Lightweight Paradox: theoretically efficient algorithms, like token shifting are currently not optimized at the low-level hardware, they do not get the same level of low-level hardware optimization (i.e. CUDA kernel tuning) that traditional convolutions have long enjoyed. In turn, the reduction in floating-point operations in consumer-grade GPUs is, therefore, overshadowed by the memory-access overhead.

4.2 The Cost of Volumetric Context

Transformer based models like Swin-UNETR dominate the literature for 3D segmentation accuracy. However, our results demonstrate that this performance comes at an prohibitive cost (85.5 GFLOPs per patch). For real-time applications, such as image-guided surgery, the inference latency ($>$ 200 ms) on mid range hardware poses a significant bottleneck, suggesting that these models are better suited to offline analysis on server scale infrastructure.

4.3 Foundation Models and Practicality

Foundation models are a paradigm shift towards zero shot generalization. However, there are models like WS -ICL, which move the burden of computations during training to inference. The need to do contextual examples processing and the target image resulted to the maximum level of computation cost of over ($>$ 1000 GFLOPs). Even though this offers potential to reduce the amount of effort needed in annotation, the available Foundation models cannot currently be easily deployed on typical clinical workstations or even mobile diagnostic devices, because they are computationally infeasible.

4.4 Deployment Recommendations

According to the efficiency-accuracy trade-off, we propose the following deployment tiers:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

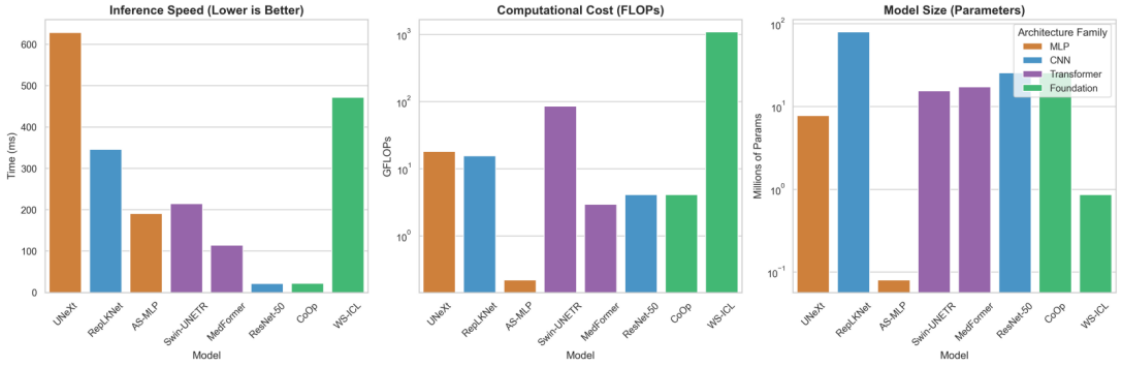

Figure 3: Comparative evaluation of computational efficiency of the eight benchmarked models. (Left) Inference Latency highlights the "Lightweight Paradox," where the MLP-based UNeXt is significantly slower than standard CNNs. (Center) Computational cost (logarithmic scale) shows the high demand for FLOPs of 3D foundation models (WS-ICL). (Right) Model Size illustrates that parameter count does not directly correlate with inference speed on standard GPU hardware.

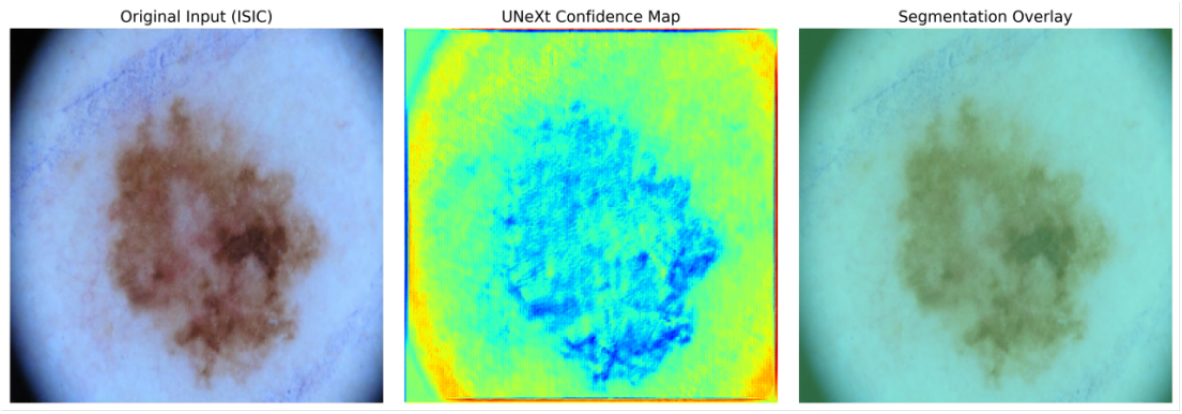

Figure 4: Qualitative segmentation results UNeXt applied to samples of ISIC 2018 skin lesions

- **Tier 1 (Mobile/Edge):** AS-MLP outperforms the alternatives: it provides the lowest computational cost (0.22 GFLOPs) and under 200 ms latency. This aligns perfectly with recent pushes for fully automatic, lightweight medical segmentation models designed specifically for resource-limited regions [39].

- **Tier 2 (Clinical Workstation):** RepLKNet and ResNet-50 are the most balanced systems of conventional two-dimensional diagnostic use.

- **Tier 3 (High-Performance Cluster):** Swin-UNETR will only be used to perform volumetric analyses in cases where the need of utmost accuracy overwhelms the limitation of computation.

5 Conclusion

This research paper fills the gap between theoretical design of architecture and practical clinical implementation. Through the benchmarking of eight representative models (two from the CNN family, two from the Transformer family, two from the MLP family, and two Foundation Models), on a standardized NVIDIA RTX3050, we are provided with realistic evaluation of computational efficiency.

Our results discard the widely accepted belief that a decrease in the number of parameters will inevitably result in a faster execution speed, by stating the unoptimization of MLP operations in the UNeXt architecture. Furthermore, we quantified the massive computational overhead of Foundation Models, identifying them as a current bottleneck for real-time systems. The next step in work will be to optimize MLP kernels with edge devices and explore hybrid CNN-Transformer models to balance on a global scale and speed on a local scale.

Acknowledgements

The authors wish to state both their appreciation and acknowledgement of the fact that the Department of Software Engineering at Delhi Technological University provided an academic environment that was favorable

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

to this research. Besides, we give credit to the open-source research community for contributing to publicly accessible implementations of the UNeXt and Swin-UNETR models.

References

[1] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] N. Tajbakhsh et al., "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, 2020.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, pp. 234–241, 2015.

[4] X. Ding et al., "Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs," *CVPR*, pp. 11963–11975, 2022.

[5] F. Isensee et al., "nnU-Net: Self-configuring Framework for Deep Learning-based Biomedical Segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.

[6] Z. Zhou et al., "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE TMI*, vol. 39, no. 6, pp. 1856–1867, 2020.

[7] O. Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," *MIDL*, 2018.

[8] H. Huang et al., "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," *ICASSP*, pp. 1055–1059, 2020.

[9] Z. Liu et al., "A ConvNet for the 2020s," *CVPR*, pp. 11976–11987, 2022.

[10] A. Hatamizadeh et al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors," *MICCAI Brainlesion Workshop*, 2022.

[11] Y. Xia et al., "MedFormer: A Hierarchical Medical Vision Transformer with Dual Sparse Selection Attention," *MICCAI*, 2023.

[12] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv:2102.04306*, 2021.

[13] H. Cao et al., "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," *ECCV Workshops*, 2022.

[14] W. Wang et al., "TransBTS: Multimodal Brain Tumor Segmentation Using Transformer," *MICCAI*, 2021.

[15] A. Hatamizadeh et al., "UNETR: Transformers for 3D Medical Segmentation," *WACV*, pp. 574–584, 2022.

[16] W. Wang and A. Howard, "MOSAIC: Mobile Segmentation via decoding Aggregated Information and Context," *arXiv:2112.11623*, 2021.

[17] D. Lian et al., "AS-MLP: An Axial Shifted MLP Architecture for Vision," *ICLR*, 2022.

[18] J. M. J. Valanarasu and V. M. Patel, "UNeXt: MLP-based Rapid Medical Image Segmentation Network," *MICCAI*, 2022.

[19] S. Chen et al., "CycleMLP: A MLP-like Architecture for Dense Prediction," *ICLR*, 2022.

[20] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *Journal of Medical Imaging*, vol. 6, no. 1, p. 014006, 2019. DOI: 10.1117/1.JMI.6.1.014006.

[21] J. Li et al., "ConvMLP: Hierarchical Convolutional MLPs for Vision," *arXiv:2109.04454*, 2021.

[22] K. Zhou et al., "Learning to Prompt for Vision-Language Models," *IJCV*, vol. 130, pp. 2337–2348, 2022.

[23] J. Hu et al., "Weakly Supervised In-Context Learning for Medical Image Segmentation," *NeurIPS*, 2024.

[24] A. Kirillov et al., "Segment Anything," *arXiv:2304.02643*, 2023.

[25] S. He et al., "Segment Anything (SAM) – Medical Benchmark Study," *arXiv:2304.09324*, 2023.

[26] M. Gaillochet et al., "Prompt learning with bounding box constraints for medical image segmentation," *arXiv preprint*, arXiv:2507.02743, 2025.

[27] W. Suo et al., "SCS: Stepwise Context Search for In-Context Segmentation," *arXiv:2407.10233*, 2024.

[28] J. Zhu et al., "MedSAM-2: A Generalized Medical Image Segmentation Foundation Model," *arXiv:2408.00874*, 2024.

[29] K. Zhou et al., "Conditional Prompt Learning for Vision-Language Models," *CVPR*, pp. 16816–16825, 2022.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CVPR*, pp. 770–778, 2016.

[31] M. E. Rayed et al., "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, p. 101504, 2024.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

[32] Y. Xu et al., "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering* (MDPI), vol. 11, no. 10, p. 1034, 2024.

[33] P. Liang et al., "Current and emerging trends in medical image segmentation with deep learning," *IEEE Access*, vol. 11, 2023.

[34] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[35] Ö. Çiçek et al., "3D U-Net: Learning dense volumetric segmentation from sparse annotation," *MICCAI*, pp. 424–432, 2016.

[36] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[37] J. Wu et al., "Medical SAM adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint*, arXiv:2304.12620, 2023.

[38] Y. Tang et al., "Self-supervised pre-training of Swin Transformers for 3D medical image analysis," *CVPR*, pp. 20730–20740, 2022.

[39] J. Wu et al., "Embed-MedSAM: A fully automatic lightweight medical segmentation model for resource-limited regions," *npj Digital Medicine*, 2025.

[40] A. Khan et al., "A recent survey of vision transformers for medical image segmentation," *arXiv preprint*, arXiv:2312.05566, 2023.