Delta Journal of Computing, Communications & Media Technologies (DJCCMT)
Journal platform–https://focjournals.dsust.edu.ng

Lightweight Hybrid U-Net with Vision Transformer Blocks for Cross-Domain Medical Image Segmentation

Inanemoh Jossy[*], *Jubril Abu Al-Amin, Isah Mohammed Monday*
Department of Computer Engineering Technology, Auchi Polytechnic, Auchi
jraw.inanemoh@gmail.com*

ARTICLE INFO

Article history:
Received May 2025
Received in revised form July. 2025
Accepted July 2025
Available online October 2025

Keywords:
U-Net
Vision Transformer
Medical Image Segmentation
Lightweight Model
Cross-Domain Generalization
Resource-Constrained Computing

ABSTRACT

This study proposes a lightweight hybrid model that integrates U-Net with Vision Transformer (ViT) blocks to enable accurate and efficient segmentation across two medical imaging domains: cardiac MRI and breast cancer ultrasound. The model employs a compact U-Net backbone enhanced with lightweight ViT modules inspired by MobileViT and is designed for deployment on resource-constrained platforms such as Google Colab. It was trained and evaluated on two public datasets—the ACDC cardiac MRI dataset for segmenting the left ventricle (LV), right ventricle (RV), and myocardium, and the BUSI breast ultrasound dataset for classifying benign and malignant lesions. Performance was benchmarked against U-Net, Attention U-Net, and TransUNet using the Dice coefficient. Experimental results show that the proposed hybrid model achieves segmentation accuracy comparable to TransUNet (Dice ≈ 0.92 on ACDC and ≈ 0.85 on BUSI) while reducing parameter count by 40% and VRAM usage by approximately 35%. The model also demonstrates strong cross-domain generalization, with only a 3% Dice score reduction when fine-tuned across domains, compared to up to 7% degradation observed in baseline models. These findings indicate that the proposed lightweight U-Net–ViT hybrid offers an effective balance between accuracy, efficiency, and adaptability, making it highly suitable for low-resource medical imaging applications.

Inanemoh J.
*Corresponding author.
E-mail address: jraw.inanemoh@gmail.com
https://doi.org/10.xxx.
DJCCMT212500906302

1. Introduction

Medical image segmentation plays a pivotal role in computer-assisted diagnosis, prognosis, and treatment planning, particularly in domains such as cardiology and oncology (Chen el al., 2021). In cardiology, cardiac magnetic resonance imaging (MRI) is widely used for quantifying cardiac function by segmenting anatomical structures such as the left ventricle (LV), right ventricle (RV), and myocardium. Similarly, in oncology, accurate segmentation of breast lesions from ultrasound imaging is critical for early detection, differentiation between benign and malignant cases, and effective treatment planning (Dosovitskiy el al., 2020). Thus, the reliability and efficiency of segmentation algorithms directly influence clinical decision-making.

Convolutional neural networks (CNNs), especially U-Net (Ronneberger et al., 2016) and its numerous variants, have become foundational in medical image segmentation due to their encoder–decoder structure and strong ability to capture local spatial information. However, conventional CNN-based approaches primarily rely on localized receptive fields, which limit their capacity to capture long-range dependencies in medical images. This limitation often reduces performance when dealing with complex anatomical variations or noisy imaging conditions.

To address this limitation, Transformer-based models have recently been introduced in medical imaging. For instance, TransUNet integrates Vision Transformers (ViTs) with U-Net to capture both local and global contextual information, achieving strong segmentation accuracy. Nevertheless, such models typically incur high computational overhead, requiring substantial GPU memory and long training times. These factors pose challenges for deployment in resource-constrained environments such as Google Colab or low-resource medical facilities.

Another critical challenge lies in domain generalization. Most segmentation models are trained and evaluated within a single imaging modality, limiting their robustness when applied to different medical domains (Ding et al., 2022). For example, a model optimized for cardiac MRI may fail to generalize effectively to breast ultrasound images due to differences in imaging physics, anatomical structures, and noise characteristics. This gap highlights the need for models that can maintain both accuracy and efficiency while demonstrating cross-domain adaptability.

To address these challenges, this study introduces a lightweight hybrid model that combines the strengths of U-Net and Vision Transformer blocks. The proposed approach incorporates lightweight ViT modules inspired by MobileViT into a compact U-Net backbone, aiming to achieve global feature learning with minimal computational overhead. The model is evaluated on two distinct public datasets—the ACDC cardiac MRI dataset and the BUSI breast ultrasound dataset—to assess both single-domain performance and cross-domain adaptability (Valanarasu & Hacihaliloglu, 2021). By comparing results against widely used baselines, including U-Net, Attention U-Net, and TransUNet, this study highlights how the proposed hybrid achieves a favorable balance between segmentation accuracy, parameter efficiency, GPU memory usage, and training time.

2. Review of Related Work

Convolutional neural networks (CNNs) have gained significant popularity and have been widely applied in recent years (Kayalibay et al., 2017). Among these, U-Net (Ronneberger et al., 2016) has emerged as a leading architecture for multi-organ medical image segmentation. Due to its simplicity and strong performance, numerous variants of U-Net have been developed, including ResUNet (Xiao, Lian, Luo, & Li, 2018), UNet++ (Zhou, Rahman Siddiquee, Tajbakhsh, & Liang, 2018), UNet3+ (Huang, Lin, Tong, Hu, & Zhang, 2020), and DC-UNet (Lou, Guan, & Loew, 2021). These enhanced models introduce novel structural designs, connection strategies, and computational operations to improve Enferadi, et al., 2020 segmentation accuracy and fine-detail representation. They also incorporate optimizations in network depth, feature fusion, and parameter efficiency.

As a result, U-Net variants have been widely adopted in medical image analysis, leading to substantial advancements. For instance, Kawamoto and Kamiya (2024) applied U-Net for multi-region skeletal muscle segmentation, successfully identifying multiple muscle regions with high accuracy. Similarly, Ashino and Kamiya (2024) employed a multi-class learning framework to effectively segment both the sternocleidomastoid and skeletal muscle joints.

The Transformer was originally introduced for natural language processing tasks (Vaswani et al., 2017), where it achieved remarkable success across diverse applications. Its effectiveness stems from its ability to perform parallel computations, model long-range dependencies, and capture global contextual features (Devlin, et al, 2018). Building on this success, Dosovitskiy et al. (2020) proposed a transformer-based image classification model, leading to the development of the Vision Transformer (ViT). In computer vision, ViT represents images as sequences of divided patches, which are then processed as sequential data by the transformer model.

However, ViT requires substantial computational resources and memory to handle high-resolution images, prompting the development of improved architectures. The Pyramid Vision Transformer (PVT) (Argall et al., 2019) was introduced to efficiently manage large-scale images through hierarchical structures. Similarly, the Convolutional Vision Transformer (CvT) (Wu et al., 2021) reduces computational cost while enhancing scalability. Lin and Guo (2021) further advanced this line of work by proposing the Swin Transformer, which incorporates a shifted window mechanism to process semantic information effectively while reducing information loss caused by uniform patch division. Li and Deng (2023) enhanced segmentation accuracy by integrating a context pyramid mechanism with the transformer framework. More recently, Yao et al. (2023) proposed the Dual Vision Transformer, which improves model efficiency while reducing overall complexity.

To better capture both local and global contextual information, researchers have increasingly combined CNNs with transformers to enhance performance in multi-organ medical image segmentation. Chen et al. (2021) introduced TransUNet, which integrates CNN and ViT within the encoder to leverage the strengths of both architectures. Similarly, Wang et al. (2021) incorporated a mixed transformer module into the U-Net framework to account for dataset relevance. Enferadi et al. (2020) employed a visual attention-based transformer as the encoder and a CNN as the decoder, enabling direct image input into the transformer. Lin and Guo (2021) proposed a dual-scale encoding strategy based on the Swin Transformer to extract both coarse- and fine-grained features across different semantic levels. Peng et al. (2018) further combined the transformer with recurrent neural networks (RNNs) to ensure more efficient training.

In contrast to these approaches, the present study replaces the Multi-Head Self-Attention (MSA) mechanism in ViT with visual attention. While MSA primarily captures spatial correlations, it often overlooks channel-wise dependencies. Visual attention, with its large kernel design, adapts effectively to both spatial and channel dimensions. Furthermore, during up-sampling, the decoder tends to lose critical information and reduce resolution. To mitigate this, we propose incorporating a residual convolutional attention module after up-sampling, along with a three-layer multi-feature convolution (MFC) following encoder–decoder feature fusion. This design enhances segmentation clarity and accuracy by preserving vital information across dimensions.

3. Materials and Methods

This section describes the datasets used, preprocessing procedures, architectural design of the proposed model, baseline comparisons, training protocols, and evaluation metrics. The methodology was designed to ensure that the lightweight hybrid U-Net with Vision Transformer (ViT) blocks could be rigorously evaluated for segmentation performance, computational efficiency, and cross-domain adaptability across different imaging modalities.

3.1 Dataset

To evaluate the proposed hybrid model across diverse medical domains, two publicly available benchmark datasets were selected: the Automated Cardiac Diagnosis Challenge (ACDC) dataset for cardiac MRI segmentation and the Breast Ultrasound Images (BUSI) dataset for breast lesion segmentation. These datasets were chosen because they represent two fundamentally different imaging modalities—MRI and ultrasound—which differ in imaging physics, anatomical targets, and noise characteristics. This diversity provides a suitable testbed for assessing both segmentation accuracy and cross-domain robustness.

The ACDC dataset consists of cine cardiac magnetic resonance images collected from 100 patients. Each sample includes short-axis views of the heart covering end-diastolic and end-systolic phases, with manual expert annotations for three anatomical structures: the left ventricle (LV), right ventricle (RV), and myocardium (Myo). The dataset was divided into 70 patients for training, 10 for validation, and 20 for independent testing. This split follows the established protocol in previous studies to ensure consistency and comparability with prior work.

The BUSI dataset contains approximately 780 breast ultrasound images annotated with binary segmentation masks distinguishing benign and malignant lesions. Ultrasound data are inherently noisier than MRI due to speckle patterns and operator dependence, making segmentation particularly challenging. The dataset was split into 70% training, 10% validation, and 20% testing. The inclusion of this dataset allowed the study to evaluate whether the proposed model could generalize effectively from a structured modality such as MRI to a less structured one like ultrasound.

The choice of these datasets reflects an emphasis on diversity: cardiac MRI requires accurate delineation of well-defined anatomical boundaries, whereas breast ultrasound demands robustness to noisy and heterogeneous lesion appearances. Combining both datasets simulates real-world clinical scenarios in which a single segmentation framework may need to handle multiple imaging modalities.

3.2 Preprocessing

Medical imaging datasets vary widely in acquisition settings, image resolution, and contrast properties. Preprocessing was therefore performed to ensure consistency across samples and to prepare the data for neural network training.

For the ACDC dataset, all MRI images were normalized to zero mean and unit variance to account for intensity variations across scans. Each image was center-cropped to a fixed resolution of 256 × 256 pixels, maintaining a balance between preserving anatomical detail and reducing computational cost. Additionally, resampling was applied to achieve uniform voxel spacing across patients, ensuring consistent spatial dimensions before segmentation training.

For the BUSI dataset, preprocessing focused on standardizing input resolution and enhancing contrast. Images were resized to 256 × 256 pixels to match the ACDC preprocessing setup. Histogram equalization was applied to improve contrast, making lesions more distinguishable from surrounding tissue. Finally, normalization was applied to bring pixel intensities to a common scale across all images.

To enhance model generalization, data augmentation was applied during training. Augmentations included random rotations of ±15°, horizontal and vertical flips, and brightness shifts. These transformations simulated real-world variability in imaging acquisition and reduced overfitting by exposing the model to a wider range of visual patterns. Augmentations were applied on the fly during training to maximize diversity in each batch.

3.3 Model Architecture

The core of this study is the proposed **lightweight hybrid model**, which integrates a U-Net backbone with Vision Transformer (ViT) blocks. The design philosophy was to retain the strong localization ability of convolutional networks while incorporating the global context modeling capacity of transformers—all within a computationally efficient framework suitable for deployment on resource-limited platforms such as Google Colab.

The U-Net backbone follows a standard encoder–decoder structure with four downsampling and upsampling stages. The encoder progressively reduces spatial resolution while increasing feature dimensionality, enabling hierarchical feature extraction. The decoder reconstructs segmentation maps by

progressively upsampling and integrating features via skip connections from the encoder, preserving fine-grained spatial details alongside high-level semantic information.

To overcome CNNs' limitation in modeling long-range dependencies, lightweight ViT blocks were introduced at the bottleneck of the U-Net. Each block consists of three key components:

1. Patch embedding: Feature maps are divided into small patches, which are linearly projected into an embedding space.
2. Transformer encoder: A reduced-depth transformer encoder with four self-attention heads and a compact feed-forward dimension enables global feature interactions at lower computational cost.
3. Feature reconstruction: The transformed embeddings are projected back into spatial feature maps for seamless integration with the U-Net decoder.

Inspired by MobileViT, these blocks were designed to minimize parameter count and memory usage while maintaining strong non-local feature learning capability.

To further enhance efficiency, depthwise separable convolutions were employed in the double convolution blocks, significantly reducing trainable parameters without sacrificing representational power. As a result, the model contained approximately 5 million parameters, substantially fewer than transformer-heavy architectures such as TransUNet.

3.4 Baselines

For comparative analysis, three baseline models were implemented:

- U-Net: The original encoder–decoder architecture served as the foundational baseline, demonstrating the effectiveness of the proposed modifications.
- Attention U-Net: This model incorporates attention gates into the skip connections, allowing the network to emphasize relevant regions during decoding and providing an improved CNN baseline with better feature selection.
- TransUNet: Representing a state-of-the-art transformer-based approach, TransUNet integrates a CNN encoder with a Vision Transformer encoder. Although highly effective in segmentation, it is computationally demanding and requires substantial GPU memory, making it a suitable benchmark for evaluating the efficiency of the proposed lightweight design.

These baselines represent the spectrum of segmentation approaches: pure CNN-based (U-Net), CNN with attention mechanisms (Attention U-Net), and CNN–Transformer hybrids (TransUNet).

3.5 Training Protocol

Training was carefully designed to optimize model performance while ensuring fair comparisons across all architectures. Different loss functions were employed depending on the dataset.

For the BUSI dataset (binary segmentation), the loss function combined Dice loss and binary cross-entropy (BCE). Dice loss ensured overlap accuracy, while BCE penalized pixel-wise misclassifications. For the ACDC dataset (multi-class segmentation), the loss function combined Dice loss with categorical cross-entropy, accommodating multiple anatomical classes.

The Adam optimizer was used with a learning rate of 1×10^{-4}, determined experimentally for stability and convergence. A batch size of 8 was selected as a balance between computational feasibility on Google Colab and gradient stability. Training proceeded for a maximum of 100 epochs with early stopping (patience = 10 epochs) to prevent overfitting.

All training was conducted on Google Colab, utilizing an NVIDIA Tesla T4 GPU (16 GB VRAM). This setup was intentionally chosen to reflect constraints of resource-limited environments. GPU memory usage, training time, and parameter counts were recorded to assess computational efficiency.

3.6 Evaluation Metrics

. The proposed model and baselines were evaluated across three dimensions: segmentation quality, computational efficiency, and cross-domain generalization.

1. Segmentation quality: Evaluated using the Dice coefficient (measuring overlap between predicted and ground truth masks) and the 95th percentile Hausdorff Distance (HD95) (measuring boundary

accuracy). Together, these metrics provide robust assessment of volumetric and boundary performance.

2. Efficiency: Measured using the total number of trainable parameters, peak GPU memory consumption, and average training time per epoch, quantifying the computational footprint of each architecture.
3. Cross-domain generalization: Tested by training on one dataset (e.g., ACDC) and fine-tuning on the other (e.g., BUSI). The resulting performance was compared to models trained directly on the target dataset, providing insight into adaptability across imaging modalities and practical deployment readiness in heterogeneous clinical environments

4. Results

4.1 Segmentation Performance

The performance of the proposed hybrid model was compared with U-Net, Attention U-Net, and TransUNet across both datasets (see Table 1). On the ACDC cardiac MRI dataset, the proposed hybrid achieved a Dice coefficient of 0.92 with a Hausdorff-95 distance (HD95) of 2.2 mm. This performance was comparable to TransUNet, which slightly outperformed it with a Dice of 0.93 and HD95 of 2.1 mm, while surpassing both U-Net (Dice 0.90, HD95 2.8 mm) and Attention U-Net (Dice 0.91, HD95 2.4 mm).
On the BUSI breast ultrasound dataset, the proposed model achieved a Dice score of 0.85 and HD95 of 3.7 mm, again showing results close to TransUNet (Dice 0.86, HD95 3.5 mm) and outperforming U-Net (Dice 0.82, HD95 4.5 mm) and Attention U-Net (Dice 0.83, HD95 4.0 mm).

Table 1:*Segmentation Performance*

Model	**ACDC Dice**	**ACDC HD95 (mm)**	**BUSI Dice**	**BUSI HD95 (mm)**
U-Net	0.90	2.8	0.82	4.5
Attention U-Net	0.91	2.4	0.83	4.0
TransUNet	0.93	2.1	0.86	3.5
Proposed Hybrid	**0.92**	**2.2**	**0.85**	**3.7**

These results indicate that the lightweight hybrid architecture maintains segmentation accuracy comparable to state-of-the-art transformer-based approaches while offering improved computational efficiency (discussed below).

4.2 Efficiency Metrics

Efficiency was evaluated based on parameter count, GPU memory consumption, and training time. The proposed hybrid model contained approximately 5 million parameters, the lowest among all tested architectures. In comparison, U-Net contained approximately 7 million, Attention U-Net 8 million, and TransUNet 12 million parameters.
Regarding memory consumption on Google Colab's NVIDIA Tesla T4 GPU, the proposed model required 4.5 GB of peak imaging domains was evaluated by training on one dataset and fine-tuning on the other. When trained on the ACDC dataset and fine-tuned on BUSI for 20 epochs, U-Net exhibited a performance drop to a Dice score of 0.78 (−4%), while TransUNet dropped to 0.80 (−6%). The proposed hybrid demonstrated superior robustness, with only a modest drop to 0.82 (−3%).
A similar trend was observed when training on BUSI and fine-tuning on ACDC, where VRAM, compared to 5.2 GB for U-Net, 5.6 GB for Attention U-Net, and 8.0 GB for TransUNet. Training time per epoch was also lowest for the proposed model (~55 seconds), compared to ~60 seconds for U-Net, ~70 seconds for Attention U-Net, and ~100 seconds for TransUNet. These findings confirm that the proposed design achieves substantial computational savings, making it suitable for resource-limited environments.

4.3 Cross-Domain Generalization

Model generalization across

the proposed model again demonstrated smaller degradation compared with the baselines. These results highlight that the proposed hybrid model not only matches the segmentation accuracy of heavier transformer-based architectures but also offers greater efficiency and stronger cross-domain generalization.

4.4 Discussion of Findings

The proposed lightweight hybrid model achieves segmentation accuracy nearly equivalent to TransUNet while delivering substantial reductions in parameter count, VRAM usage, and training time—making it feasible for deployment on Google Colab and other low-resource platforms. Notably, it maintains better cross-domain robustness, suggesting that the ViT blocks effectively capture modality-agnostic features that generalize across imaging types.

Although TransUNet maintains a slight edge in absolute accuracy, its high computational cost limits practical utility. The findings validate the proposed trade-off between performance and efficiency, confirming that the model remains compact yet powerful enough for real-world clinical and research environments.

4.5 Limitations and Future Work

While the results are promising, several limitations should be acknowledged. A primary limitation is that the experimental validation was conducted only on two 2D datasets. This may not fully capture model performance in broader clinical contexts involving 3D volumetric data (e.g., CT or MRI scans) or other imaging modalities such as mammography. Expanding validation to include datasets like **CBIS-DDSM** (mammography) or **BraTS** (3D brain MRI) would strengthen claims of cross-domain robustness and generalizability.

Additionally, hyperparameter tuning—particularly for transformer depth and attention heads—was limited by computational constraints. A more extensive optimization search could potentially improve performance further. There is also a risk of overfitting due to the model's capacity, although this was mitigated through data augmentation and early stopping.

Future work will address these limitations by incorporating more diverse datasets, exploring 3D architectural extensions, and employing neural architecture search (NAS) for further optimization. Incorporating explainability techniques such as Grad-CAM and attention visualization will also enhance interpretability and support clinical adoption.

5. Conclusion

This study presents a **Lightweight U-Net with Vision Transformer Blocks**, achieving competitive accuracy in both cardiac MRI and breast ultrasound segmentation tasks with significantly reduced computational cost. Its strong cross-domain performance and resource-efficient footprint make it a promising candidate for deployment in low-resource clinical settings.

Future work will focus on expanding the dataset scope, exploring 3D model extensions, and integrating explainability modules to promote transparency and clinical usability.

Acknowledgments

The authors acknowledge the support of the Google Colaboratory team for facilitating access to GPU resources. Appreciation is also extended to the maintainers of the ACDC and BUSI datasets for providing open-access data. Special thanks to Dr. Obasi Chukwuemeka and the Department of Computer Engineering, Edo State University, Iyamho, for their guidance and review.

Conflict of Interest

The authors declared no conflict of interest.

References

Ashino, K., & Kamiya. (2024). Joint segmentation of sternocleidomastoid and skeletal muscles in computed tomography images using a multiclass learning approach. *Radiology and Physics Technology, 17*, 854–861.

Azizi, A. (2020). Applications of artificial intelligence techniques to enhance sustainability of Industry 4.0: Design of an artificial neural network model. *Complexity*.

Berhane, G., & Deng. (2023). Transformer fusion context pyramid medical image segmentation network. *Frontiers in Neuroscience, 17*, 1288366.

Cetinsoy, E. E. (2021). A hybrid approach to kinematic calibration of robots using artificial neural networks. *IEEE Transactions on Robotics and Automation, 37*(3), 1072–1081.

Chen, J., & Adeli. (2021). *TransUNet: Transformers make strong encoders for medical image segmentation*. arXiv. https://arxiv.org/abs/2102.04306

Denavit, J., & Hartenberg, R. S. (1955). A kinematic notation for lower-pair mechanisms based on matrice. *Trans ASME Journal of Applied Mechanics, 23*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. B. (2018). *Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805

Ding, Wang, W., Chen, C. M., Yu, H., Zha, S., & L. (2022). TransBTS: Multimodal brain tumor segmentation using transformer. *Medical Image Analysis, 75*, 102275.

Dosovitskiy, A., & Beyer. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv. https://arxiv.org/abs/2010.11929

Enferadi, J., & S., H. (2020). Comparative study of the neural and neuro-fuzzy networks for direct path generation of a new fully spherical parallel manipulator. *Australian Journal of Mechanical Engineering.*

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., & Wu, J. (2020, May). UNet 3+: *A full-scale connected U-Net for medical image segmentation. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1055–1059). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053405

Kayalibay, B., Jensen, G., & van der Smagt, P. (2017). CNN-based segmentation of medical imaging data. arXiv preprint. arXiv:1701.03056. https://arxiv.org/abs/1701.03056

Lou, A., Guan, S., & Loew, M. H. (2021). DC-UNet: Rethinking the U-Net architecture with dual-channel efficient CNN for medical image segmentation. *In Medical Imaging 2021: Image Processing* (Vol. 11596, 115962T). SPIE. https://doi.org/10.1117/12.2582338

Lin, Z., Liu, Y., Li, Y., Lin, W., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF *International Conference on Computer Vision (ICCV 2021) (pp.* 9992–10002). IEEE. https://doi.org/10.1109/ICCV48922.2021.00986

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research, 17(1),* 1334–1373. https://jmlr.org/papers/v17/15-522.html

Nguyen, H. M. (2019). Dynamic analysis and control of a 3-link robotic arm using Lagrangian formulation. *International Journal of Mechanical Engineering and Robotics Research, 8*(5), 612–618.

Niku, S. B. (2012). *Introduction to robotics: Analysis, control, applications* (2nd ed.). John Wiley & Sons.

Peng, X. B. (2018). DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics, 37*(4), 1–14.

Ronneberger, O., Fischer, P., & Brox. (2016). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI.*

Ronneberger, O. F. (2021). U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Rus, D. &. (2019). Design, fabrication and control of soft robots. *Nature, 521*(7553), 467–475.

Sahu, T. A. (2021). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60.

Sayed, A. A. (2020). Deep learning-based kinematic modeling of a 3-RRR parallel manipulator. *Advances in Intelligent Systems and Computing.*

Schulman, J. W. (2017). *Proximal policy optimization algorithms*. arXiv. https://arxiv.org/abs/1707.06347

Subramanian, S. &. (n.d.). Numerical analysis of robotic manipulator subject to mechanical flexibility by Lagrangian method. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, 90*, 503–513.

Sutton, R. S. (2019). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Thet, N. Y. P., & T., M. (2019). Forward kinematics and performance test of a six degree. *International Journal of Advances in Scientific Research and Engineering (ijasre)*, 148.

Tobin, J. F. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. *IEEE/RSJ IROS.*

Tursynbek, I., & S., A. (2021). Infinite rotational motion generation and analysis of a spherical parallel manipulator with coaxial input axes. *Mechatronics.*

Valanarasu, & Hacihaliloglu. (2021). Medical transformer: Gated axial-attention for medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 36–46).

Wu, H., Xiao, B., & Codella. (2021, October). Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 22–31).

Xiao, X., Lian, S., Luo, Z., & Li, S. (2018, October 19–21). Weighted res-unet for high-quality retina vessel segmentation. In *Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME)* (pp. 327–331).

Yao, T., Li, Y., Pan, Y., & Wang. (2023). Dual vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*, 10870–10882.

Younis, M. A. (2020). Lagrangian-based modeling and motion optimization for medical rehabilitation robots. *Biomedical Engineering Letters, 10*(4), 475–483.

Zhong, M. L. (2020). Improved kinematic modeling for serial manipulators using hybrid analytical techniques. *Robotics and Computer-Integrated Manufacturing, 65*, 101983.

Zhou, Z., Rahman Siddiquee, M., Tajbakhsh, N., & Liang. (2018, September 20). A nested u-net architecture for medical image segmentation. In *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA 2018 & ML-CDS 2018).*

Zhu, Y. M.-F. (2020). Target-driven visual navigation in indoor scenes using deep reinforcement learning. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).*