MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 2040-2045

Autosomal Dominant Polycystic Kidney Segmentation Based on Multi-Scale Attention Mechanisms with Swin-UNETR

Alpaslan Yavuz[1], Elif Meseci[2,3], Caner Ozcan[4*], Iclal Erdem Toslak[1], Gizem Karagol Aydinyurt[1], Heves Yaren Karakas[1]

[1]Department of Radiology, Health Science University Antalya Training and Research Hospital, Antalya, Turkiye [2] Department of Computer Engineering, Bulent Ecevit University, Zonguldak, Turkiye
[3] Department of Computer Engineering, Institute of Graduate Programs, Karabuk University, Karabuk, Turkiye
[4] Department of Software Engineering, Karabuk University, Karabuk, Turkiye
elif.meseci@beun.edu.tr; alp_yavuz@hotmail.com; canerozcan@karabuk.edu.tr; driclalerdem@yahoo.com; gizemkaragol@gmail.com; hevesykarakas@gmail.com
*Corresponding Author: canerozcan@karabuk.edu.tr
ORCID ID: 0000-0002-2854-4005

Abstract— *Autosomal dominant polycystic kidney disease (ADPKD) is a genetic disorder characterized by the formation of numerous cysts, requiring accurate and automated methods for early diagnosis and monitoring. U-Net-based deep learning (DL) models were used to improve kidney segmentation in magnetic resonance (MR) and computed tomography (CT) images. The primary goal was to achieve accurate and efficient kidney segmentation, which serves as a critical step in the diagnosis of ADPKD. The Swin-UNETR architecture, which aims to improve segmentation accuracy using a Swin Transformer-based encoder capable of effectively modeling local and global contexts, was employed. The model was trained and validated on a specialized dataset labeled by experts, and its performance was evaluated using metrics such as the Dice similarity coefficient (DSC) and loss value. The results demonstrate that the Swin-UNETR model achieves high segmentation accuracy with an average DSC exceeding 90%. The three-dimensional segmentation outputs show that the kidney boundaries are preserved with high accuracy in each slice and that the model successfully follows anatomical integrity. The multi-scale attention mechanisms of the Swin Transformer-based architecture enhance segmentation accuracy by effectively modeling both local details and global contextual relationships. This study demonstrates that the Swin-UNETR approach, which offers reliability and reproducibility in the automatic kidney segmentation process, has clinically significant potential in the imaging-based assessment of polycystic kidney disease. The proposed method is expected to increase diagnostic efficiency by reducing manual intervention and contribute to the development of advanced clinical support systems.*

Keywords— Autosomal dominant polycystic kidney disease, multi-scale attention mechanisms, segmentation, Swin-UNETR.

Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is a genetically inherited and progressive disease characterized by the progressive development of multiple cysts within the renal parenchyma. Throughout the course of the disease, the increase in the number and distribution of cysts leads to disruption of the normal anatomical structure of the kidneys and the emergence of complex morphological appearances on imaging [1,2]. This situation necessitates the accurate and reliable assessment of kidney structure using imaging methods in the diagnosis and monitoring processes of ADPKD.

Advanced imaging techniques such as magnetic resonance imaging (MRI) and computed tomography (CT) enable detailed three-dimensional examination of the kidneys and allow for objective analysis of the morphological effects of the disease [3,4]. Kidney segmentation performed on these images is considered a fundamental step in determining kidney boundaries and evaluating structural changes in a reproducible manner. However, manual segmentation procedures are limited in their applicability in clinical practice due to the high level of expertise required, the time-consuming nature of the process, and the variability between observers [5].

To overcome these limitations, deep learning-based automatic segmentation methods have been widely adopted in recent years. Convolutional neural network architectures based on U-Net and its derivatives, in particular, have yielded successful results in the segmentation of kidneys and kidney lesions [6–9]. However, these approaches mostly focus on representations based on local features and may be insufficient in modeling long-range spatial relationships in cases where cysts such as ADPKD significantly distort kidney anatomy [10].

Recent reports indicate that Transformer-based and hybrid architectures have achieved remarkable success in medical image segmentation due to their ability to model both local details and global contextual information simultaneously [11–13]. Among these architectures, Swin Transformer can effectively learn multi-scale contextual information thanks to its hierarchical structure and sliding window-based attention mechanism. The Swin-UNETR architecture combines these advantages with U-Net-like decoder structures, offering high-accuracy segmentation potential in three-dimensional medical images [14].

This study presents a Swin-UNETR-based deep learning approach for automatically segmenting kidney structures in MR and CT images of individuals diagnosed with ADPKD. The proposed method aims to consistently and accurately determine kidney boundaries despite anatomical irregularities caused by cysts. The model was trained and validated on a specialized dataset labeled by experts. This paper addresses only the problem of automatic kidney segmentation; subsequent analysis steps such as volume calculation or prognosis estimation are beyond the scope of this study.

I. RELATED WORKS

Deep learning-based kidney segmentation has become a heavily researched area in medical image analysis in recent years. The automatic segmentation of kidney anatomy in computed tomography (CT) and magnetic resonance (MR) images, in particular, has been studied in numerous works due to its potential to reduce clinical workload and standardize analysis processes. However, this problem becomes more complex in cases where cystic structures such as ADPKD significantly irregularize kidney morphology.

U-Net and its derivatives are among the most widely used architectures in medical image segmentation, thanks to their encoder–decoder structures and skip connection mechanisms. Isensee and Hein [6] demonstrated that the 3D U-Net architecture can deliver high segmentation performance with various improvements. Similarly, Zhao et al. [7] achieved successful results in kidney and kidney tumor segmentation using a multi-scale supervised 3D U-Net approach. Studies such as Lin et al. [8] and Krishnan et al. [9] also addressed kidney and cystic lesion segmentation problems using different U-Net variants.

While these approaches are effective, particularly in learning local tissue features, they can be limited in modeling long-range spatial relationships in complex anatomies where cysts such as ADPKD irregularize kidney boundaries. This is attributed to the inability of convolutional-only architectures to adequately represent global context information.

Studies conducted specifically on ADPKD have generally focused on the accurate segmentation of kidneys and cystic structures. Shin et al. [5] demonstrated that an expert-level segmentation performance can be achieved in polycystic kidney and liver volume measurements using a deep learning-based approach. Kim et al. [10] proposed a deep learning-based method for the automatic segmentation of kidneys and exophytic cysts in individuals with ADPKD. Goel et al. [3] developed a clinically usable automatic segmentation system for polycystic kidney disease in MR images.

While these studies demonstrate that automatic analysis of ADPKD images is possible, they are mostly based on CNN-based architectures. Given the density and irregular distribution of cystic structures, it has been reported that such architectures cannot consistently preserve anatomical integrity in all cases.

In recent years, Transformer-based architectures have emerged as a compelling alternative in medical image segmentation due to their ability to model long-range dependencies. Lee et al. [11] proposed the 3D UX-Net architecture, which combines large-core convolutional layers with Transformer-like hierarchical structures, and achieved competitive results in various medical segmentation tasks. Zhang et al. [12] improved performance in abdominal organ segmentation with a multi-purpose segmentation model that combines global and local information in parallel.

In the context of kidney segmentation, it has been reported that Transformer-based approaches have the potential to improve boundary consistency, particularly in complex anatomical structures. Dual-task kidney MRI segmentation studies [13] demonstrate that Transformer-based architectures can learn effective representations for both kidney and related tasks. Fan et al. [14] proposed a three-dimensional segmentation approach based on Swin Transformer, emphasizing the importance of multi-scale contextual information in medical volume segmentation.

A review of the current literature reveals that studies on automatic kidney segmentation specifically for ADPKD largely rely on CNN-based architectures, while approaches based on Transformers or Swin Transformers have been addressed only to a limited extent. Furthermore, it is noteworthy that approaches capable of modeling both local anatomical details and global contextual relationships on three-dimensional images have not been sufficiently investigated in the context of ADPKD.

This study aims to achieve consistent and highly accurate segmentation despite the anatomical irregularities caused by cysts by adapting the Swin-UNETR architecture to the kidney segmentation problem in MR and CT images of individuals with ADPKD. Thus, the study aims to bridge the gap between existing CNN-based approaches and Transformer-based methods and contribute to the literature on the automatic analysis of ADPKD images.

II. MATERIALS AND METHODS

A. Dataset and clinical annotations

In this study, imaging data from patients diagnosed with ADPKD at Antalya Training and Research Hospital were used. In order to examine the polycystic kidney structure in detail, abdominal CT and MR images of individuals with ADPKD were included in the study. These imaging modalities allow for the evaluation of the three-dimensional anatomical structure of the kidneys, enabling a detailed analysis of the effects of polycystic formations on kidney morphology. The images used consist of three-dimensional volumetric data including coronal, sagittal, and axial planes. Each volumetric image is separated into sequential two-dimensional slices, allowing for layered examination from different angles. This structure contributes to a more consistent assessment of kidney boundaries and morphological features.

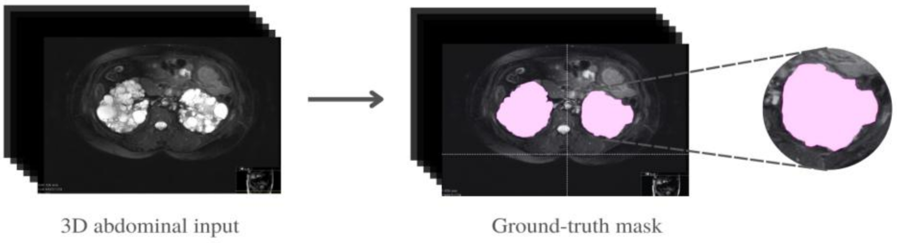

Fig. 1 Data annotation process.

A total of 141 patient image datasets were used in the study. Sequences consisting of an average of 80 slices from each patient's images were included in the labeling process. The ground truth masks used for kidney segmentation were created manually by radiology specialists. During the labeling process, kidney structures were marked in mask format using the Medical Imaging Interaction Toolkit (MITK) software as demonstrated in Fig 1. The acquired kidney masks and corresponding images were saved in NIfTI (.nii) format and anonymized to protect patient privacy. The entire dataset was used to address a challenging kidney segmentation problem such as ADPKD, where boundary uncertainties increase due to cystic structures and anatomical variations are pronounced.

B. Pre-processing

The CT and MRI images underwent a series of preprocessing and data augmentation steps before being fed into the deep learning model. First, all images and their corresponding ground truth segmentation masks were loaded and converted into single-channel volumetric data. To reduce spatial inconsistencies arising from different imaging protocols, all volumes were resampled to a fixed voxel size, and their anatomical orientations were transformed to the RAS (Right–Anterior–Superior) coordinate system.

Image intensity values were normalized, considering Hounsfield units, which are commonly used in CT images. In this context, intensity values were clipped within a specific range and scaled to the [0,1] range. Subsequently, foreground-based clipping was applied to eliminate background regions in the images and ensure the model focused on the region of interest.

During model training, fixed-size three-dimensional patches were randomly sampled from volumetric data to increase memory efficiency and make more effective use of local context information. This process was performed in a way that balanced regions containing kidneys (positive) and those without kidneys (negative).

To improve the model's generalization ability, limited data augmentation techniques were applied during the training phase. In this context, random intensity shifts, small angular rotations, and scaling operations were added to the images. All preprocessing and data augmentation steps were applied together while preserving the spatial consistency between the images and segmentation masks.

C. Deep Learning Model

In this study, the Swin-UNETR architecture based on the Swin Transformer was used to address a challenging segmentation problem where anatomical variations are prominent and organ boundaries are irregular, such as in ADPKD. Swin-UNETR is a hybrid deep learning architecture designed for three-dimensional volumetric images, aiming to effectively model global contextual relationships while preserving local details [11].

The Swin-UNETR architecture has an encoder–decoder structure. In the encoder section, hierarchical Swin Transformer blocks are used to transform the input volume into multi-scale representations. This structure enables the learning of long-range contextual relationships while keeping computational costs under control by applying a shifted window self-attention mechanism within fixed-size windows [15]. Working on three-dimensional images, this mechanism contributes to modeling the holistic structure of organs with extensive spatial coverage, such as the kidney. The model architecture is shown in Fig 2.

The multi-scale feature maps obtained from the encoder are progressively merged in the decoder section, and spatial resolution is gradually recovered. Skip connections used throughout the decoder path enhance boundary sensitivity by combining low-level spatial details with high-level semantic information. This structure enables more consistent segmentation outputs in the face of irregular boundaries and shape variations commonly encountered in polycystic kidneys.

Swin-UNETR's multi-scale attention mechanisms aim to overcome the limitations arising from the restricted receptive fields of classical convolution-based architectures by enabling the joint modeling of both local tissue features and global anatomical context. In this regard, the architecture offers a powerful alternative for improving segmentation accuracy in cases involving complex kidney morphologies such as ADPKD [11, 16].

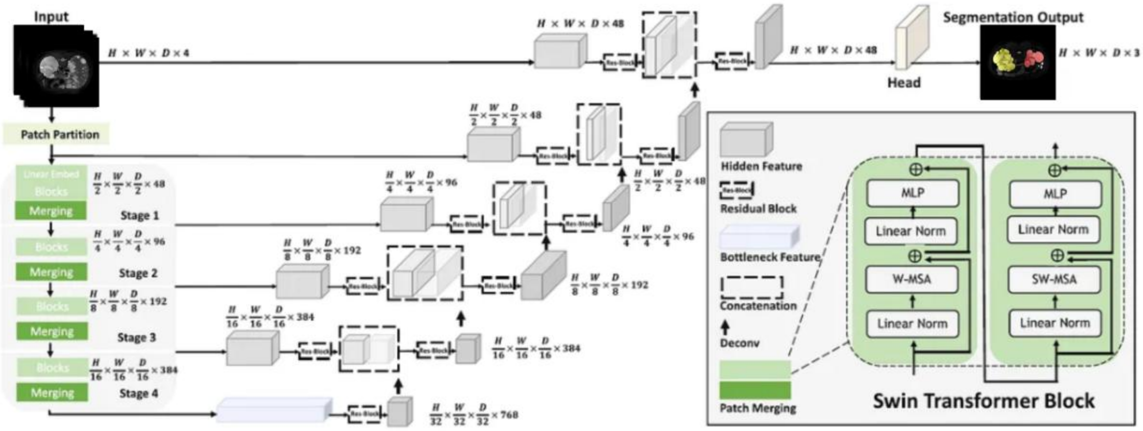
Input
$H \times W \times D \times 4$
Patch Partition
Linear Embed
Blocks
Merging
Stage 1
$\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times 48$
Blocks
Merging
Stage 2
$\frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}\times 96$
Blocks
Merging
Stage 3
$\frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}\times 192$
Blocks
Merging
Stage 4
$\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}\times 384$
$\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times 768$
$H \times W \times D \times 48$
Res-Block
Head
Segmentation Output
$H \times W \times D \times 3$
Hidden Feature
Residual Block
Bottleneck Feature
Concatenation
Deconv
Patch Merging
MLP
Linear Norm
W-MSA
SW-MSA
Swin Transformer Block

Fig. 2 Swin UNETR model architecture.

III. EXPERIMENTS AND RESULTS

A. Training Details

The proposed Swin-UNETR model was trained using the PyTorch-based MONAI library. During training, fixed-size patches randomly sampled from three-dimensional volumetric images were used. This approach was chosen to increase memory efficiency and enable the simultaneous learning of local and global context information.

The dataset was split into training and test subsets based on patients. In this context, data from 111 of the 141 patients were used for training, and data from 30 patients were used for testing. The model was trained end-to-end using the AdamW optimization algorithm. The training process was carried out over a total of 40,000 iterations, and the learning rate was kept constant. During training, a Dice-based loss function was used to improve segmentation performance. There was no overlap between patients in the training and test sets.

In the quantitative evaluation of model performance, the Dice Similarity Coefficient (DSC), commonly used in medical image segmentation studies, has been used as the primary metric. DSC enables the evaluation of the model's segmentation accuracy by measuring the overlap ratio between the predicted segmentation and the ground truth mask. Additionally, the model's convergence behaviour during the training process was monitored through loss values. All evaluation results were reported on the test dataset.

B. Quantitative results

The results obtained show that the model performs segmentation with high accuracy despite challenging anatomical variations and irregular kidney boundaries. In the evaluation performed on the test dataset, the proposed approach achieved an average DSC value of 0.9347, and the loss value calculated at the end of the training process was reported as 0.478 as shown in Table 1.

TABLE I
QUANTITATIVE RESULTS OF THE MODEL

Dataset	Iteration	DSC	Loss
Our Model	40000	0.9347	0.478
Flare	40000	0.9223	0.457

To further assess the generalization capability of the proposed model, the same architecture was trained and evaluated on the FLARE dataset. FLARE is a publicly available multi-organ abdominal CT dataset designed for benchmarking organ segmentation methods. Compared to the custom ADPKD dataset, FLARE includes anatomically normal kidneys and a broader range of abdominal structures, providing a complementary evaluation scenario for assessing model generalization. The model achieved a DSC of 0.9223 with a corresponding loss value of 0.437, demonstrating competitive segmentation performance on an external public dataset. These results indicate that the proposed approach is able to generalize beyond the custom ADPKD dataset and maintain robust performance under different data distributions and imaging characteristics.

Visualization of the expert label (ground truth) and model prediction for different slices in a case is given in Fig 3. When examining the three-dimensional segmentation outputs, it was observed that the model was able to consistently follow the kidney boundaries in all sections and preserve anatomical integrity. Even in shape distortions caused by cyst formations, it was seen that the multi-scale attention mechanisms offered by the Swin Transformer-based encoder effectively modeled both local details and global contextual information. This contributed to a significant reduction in boundary uncertainties frequently encountered in traditional convolution-based approaches.

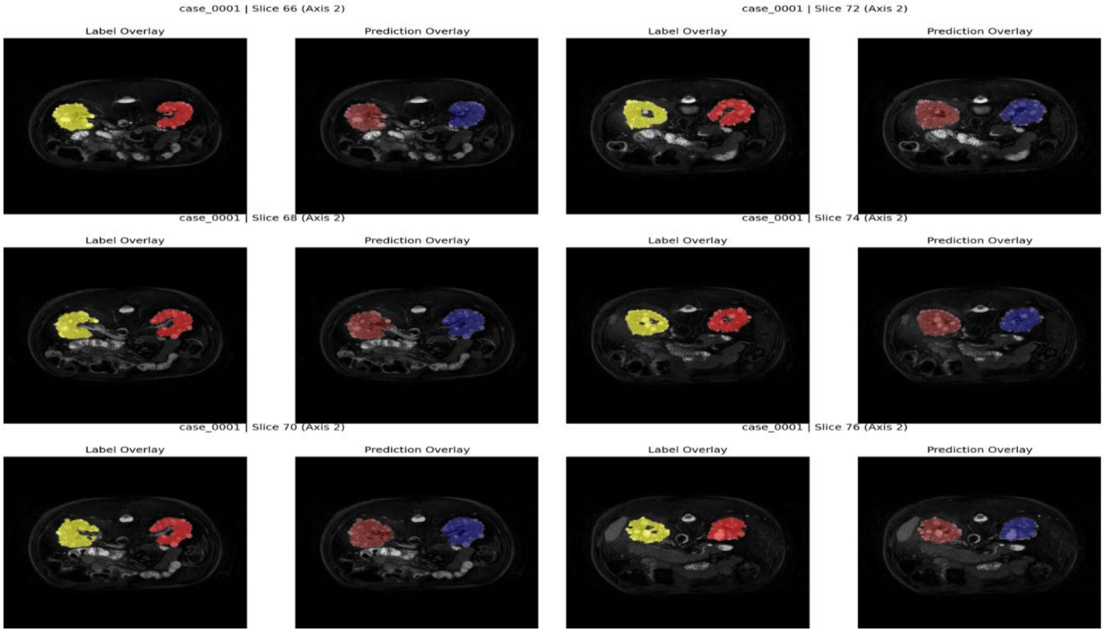

Fig. 3 Visualization of the expert label (ground truth) and model prediction for different slices in a case.

The high DSC value obtained demonstrates that the proposed method can produce reliable and reproducible results in kidney segmentation problems with complex morphologies such as ADPKD. The results show that the Swin-UNETR architecture is a strong candidate for automatic kidney segmentation on clinical images.

IV. Discussion

In this study, the Swin-UNETR architecture was used for automatic kidney segmentation in CT and MRI images of individuals diagnosed with ADPKD, yielding highly accurate results. The DSC value of 0.9347 reported on the test dataset demonstrates that the proposed approach can perform reliable segmentation despite the irregular boundaries caused by cysts and high anatomical variation.

In kidneys with ADPKD, the normal anatomical structure is significantly disrupted, and the kidney boundaries become indistinct due to heterogeneous cystic structures. This poses a significant challenge for classical convolution-based segmentation models. The results obtained reveal that the multi-scale window-based attention mechanisms offered by the Swin Transformer-based encoder can model both local boundary details and the overall geometric structure of the kidney. Thus, more consistent segmentation outputs have been obtained compared to methods based solely on local tissue features. Examination of the 3D segmentation outputs across slices demonstrates that the model preserves anatomical continuity and can produce consistent boundaries between adjacent slices. This feature is critically important for accurately modeling the contextual relationship between consecutive slices, particularly in 3D volumetric data.

In the literature, 3D U-Net and its derivatives are commonly used for the segmentation of kidneys and kidney lesions [6, 10, 15, 17]. Although studies based on nnU-Net and multi-scale 3D U-Net have achieved successful results on standard datasets, performance drops have been reported in morphologically complex diseases such as ADPKD [5, 10]. Especially in cases where cysts irregularize kidney boundaries, purely convolution-based approaches fail to adequately capture the global context. With the recent adoption of Transformer-based architectures in medical image segmentation, it has been demonstrated that long-range dependencies can be modeled more effectively. Swin-UNETR combines a hierarchical Transformer structure with a U-Net-like decoder, ensuring that both the global context and local details are learned in a balanced manner. The high DSC value obtained in this study supports that Transformer-based approaches offer a significant advantage in challenging clinical scenarios such as ADPKD.

The automatic and highly accurate kidney segmentation offered by the proposed method has the potential to reduce the need for manual segmentation in clinical imaging workflows. Considering that manual segmentation processes are time-consuming and subject to observer variability, such automatic approaches can be considered an important component for clinical decision support systems. This study focuses solely on the kidney segmentation problem, excluding advanced operations such as separate segmentation of cysts or volumetric analysis. This choice aims to examine the basic segmentation performance of the proposed method in isolation and to clearly demonstrate the effectiveness of the architecture in this challenging problem.

V. Conclusions

In this study, a deep learning approach based on Swin-UNETR is presented for ADPKD. Despite the irregular and complex kidney morphology caused by cyst formation, the proposed method was able to produce highly accurate and consistent segmentation outputs on three-dimensional images. An important aspect of this study is that it was performed on a single-center, specially created clinical dataset consisting of a limited number of patients. Unlike large-scale, publicly available datasets, this dataset reflects the high anatomical variation and advanced cystic deformities encountered in real clinical settings. The high segmentation accuracy achieved under such data-constrained conditions demonstrates the robustness of the proposed approach and its adaptability to practical clinical applications.

This study focused solely on the problem of automatic kidney segmentation, excluding advanced processes such as separate cyst segmentation or volumetric analysis from the evaluation. This approach allowed for an isolated examination of the basic segmentation capabilities of the proposed architecture. Future studies aim to produce more detailed outputs in the field of ADPKD and integrate segmentation outputs into clinical decision support systems.

Acknowledgement

This study was supported by the Health Institutes of Turkiye (TUSEB) under project number 33865.

References

[1] K. T. Bae et al., “Expanded imaging classification of autosomal dominant polycystic kidney disease,” Journal of the American Society of Nephrology, vol. 31, no. 7, 2020, doi: 10.1681/ASN.2019101121.

[2] Y. Kim et al., “Deep Learning–Based Automated Imaging Classification of ADPKD,” Kidney Int Rep, vol. 9, no. 6, 2024, doi: 10.1016/j.ekir.2024.04.002.

[3] A. Goel et al., “Deployed Deep Learning Kidney Segmentation for Polycystic Kidney Disease MRI,” Radiol Artif Intell, vol. 4, no. 2, 2022, doi: 10.1148/RYAI.210205.

[4] L. Aronson, R. Ngnitewe Massa’a, Md, S. Jamal, S. Gardezi, and A. L. Wentland, “Automatic Segmentation of the Kidneys and Cystic Renal Lesions on Non-Contrast CT Using a Convolutional Neural Network,” arXiv:2405.08282 (eess), 2024.

[5] T. Y. Shin et al., “Expert-level segmentation using deep learning for volumetry of polycystic kidney and liver,” Investig Clin Urol, vol. 61, no. 6, 2020, doi: 10.4111/icu.20200086.

[6] F. Isensee and K. H. Maier-Hein, “An attempt at beating the 3D U-Net,” 2019. doi: 10.24926/548719.001.

[7] W. Zhao and Z. Zeng, “Multi Scale Supervised 3D U-Net for Kidney and Tumor Segmentation,” 2019. doi: 10.24926/548719.007.

[8] Z. Lin et al., "Automated segmentation of kidney and renal mass and automated detection of renal mass in CT urography using 3D U-Net-based deep convolutional neural network," Eur Radiol, vol. 31, no. 7, 2021, doi: 10.1007/s00330-020-07608-9.

[9] C. Krishnan, E. Schmidt, E. Onuoha, M. Mrug, C. E. Cardenas, and H. Kim, "UNet++ Compression Techniques for Kidney and Cyst Segmentation in Autosomal Dominant Polycystic Kidney Disease," Advanced Biomedical Engineering, vol. 13, 2024, doi: 10.14326/abe.13.134.

[10] Y. Kim, C. Tao, H. Kim, G. Y. Oh, J. Ko, and K. T. Bae, "A Deep Learning Approach for Automated Segmentation of Kidneys and Exophytic Cysts in Individuals with Autosomal Dominant Polycystic Kidney Disease," Journal of the American Society of Nephrology, vol. 33, no. 8, 2022, doi: 10.1681/ASN.2021111400.

[11] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3D UX-NET: A LARGE KERNEL VOLUMETRIC CONVNET MODERNIZING HIERARCHICAL TRANSFORMER FOR MEDICAL IMAGE SEGMENTATION," in 11th International Conference on Learning Representations, ICLR 2023, 2023.

[12] G. D. Zhang et al., "MOTC: Abdominal Multi-objective Segmentation Model with Parallel Fusion of Global and Local Information," Journal of Imaging Informatics in Medicine, vol. 37, no. 3, 2024, doi: 10.1007/s10278-024-00978-2.

[13] P. H. Conze, G. Andrade-Miranda, Y. Le Meur, E. Cornec-Le Gall, and F. Rousseau, "Dual-task kidney MR segmentation with transformers in autosomal-dominant polycystic kidney disease," Computerized Medical Imaging and Graphics, vol. 113, 2024, doi: 10.1016/j.compmedimag.2024.102349.

[14] L. Fan, X. Ding, Z. Wang, H. Wang, and R. Zhang, "Medical Image Segmentation Based on 3D PDC with Swin Transformer ⋆." [Online]. Available: https://ssrn.com/abstract=5132443

[15] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proceedings of the IEEE International Conference on Computer Vision, 2021. doi: 10.1109/ICCV48922.2021.00986.

[16] X. Wang et al., "Enhanced choroid plexus segmentation with 3D UX-Net and its association with disease progression in multiple sclerosis," Mult Scler Relat Disord, vol. 88, 2024, doi: 10.1016/j.msard.2024.105750.

[17] W. Zhao, D. Jiang, J. Peña Queralta, and T. Westerlund, "MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net," Inform Med Unlocked, vol. 19, 2020, doi: 10.1016/j.imu.2020.100357.