Franklin Open 9 (2024) 100182

Contents lists available at ScienceDirect

Franklin Open

journal homepage: www.elsevier.com/locate/fraope

Temporal dependency modeling for improved medical image segmentation: The R-UNet perspective

Ahmed Alweshah [a], Roohollah Barzamini [b], Farshid Hajati [c], Shoorangiz Shams Shamsabad Farahani [d], Mohammad Arabian [b], Behnaz Sohani [e,*]

[a] Department of Electrical Engineering, Islamic Azad University Science and Research Branch, Tehran, Iran
[b] Department of Electrical Engineering, Central Tehran Branch, Islamic Azad University, Tehran 1148963537, Iran
[c] School of Science and Technology, Faculty of Science, Agriculture, Business and Law, University of New England, Armidale, NSW 2350, Australia
[d] Department of Electrical Engineering, Islamshahr branch, Islamic Azad University, Islamshahr, Iran
[e] Wolfson School of Mechanical, Electrical & Manufacturing Engineering, Loughborough University, Loughborough, Leicestershire, LE11 3TU, UK

ARTICLE INFO

Keywords:
Machine learning
Medical image processing
R-UNet
Deep learning
Recurrent neural network
CNN
Segmentation
UNet

ABSTRACT

In this study, we propose a modified version of the widely used UNet architecture, enhanced by the integration of recurrent blocks at each step of the encoder (down-sampling) and decoder (up-sampling) stages. The proposed Recurrent UNet (R-UNet) architecture aims to improve the performance of semantic segmentation tasks by allowing the model to capture temporal dependencies and long-range contextual information. The R-UNet architecture consists of two main components: a recurrent encoder and a recurrent decoder. The recurrent encoder is composed of a series of convolutional and recurrent blocks, which extract features from the input image and propagate them across time. The recurrent decoder consists of a similar series of convolutional and recurrent blocks, which use the extracted features to generate the final segmentation mask. An attention mechanism is employed to enhance feature extraction at the bottleneck of the model. The proposed R-UNet architecture is evaluated on multiple benchmark datasets, including those for liver segmentation, brain tumor detection, mitochondria segmentation, lung imaging, a proprietary lung CT COVID-19 dataset, as well as various multi-organ imaging datasets. The experimental results demonstrate that the proposed R-UNet architecture outperforms the standard UNet architecture and several other state-of-the-art semantic segmentation models in terms of accuracy score, achieving an overall accuracy of 97.2 % on the Mitochondria dataset, 97.83 % on the Liver dataset, 89.17 % on the Tumor dataset and 97.22 % Lung dataset.

1. Introduction

Medical image segmentation is a vital area of research in biomedical image analysis, as it plays a crucial role in enhancing diagnosis and treatment for various diseases. Despite the considerable progress in deep learning methodologies, existing segmentation techniques, particularly those employing meta-learning strategies, encounter significant challenges [1]. For example, many traditional models struggle with generalization across diverse datasets, leading to diminished performance in real-world clinical scenarios. Additionally, the reliance on large, well-annotated datasets presents a barrier to the practical application of these techniques, especially in environments where access to labelled data is limited [2]. This highlights a pressing need for innovative approaches that can effectively address these shortcomings while ensuring robust performance across different imaging modalities and conditions. In light of these challenges, our research introduces the R-UNet architecture, designed to overcome the limitations of conventional segmentation models. By integrating recurrent connections and attention mechanisms, R-UNet enhances the capacity to capture long-range dependencies within medical images, a critical factor given that anatomical structures can exhibit variability over time or across different imaging perspectives. The inclusion of attention mechanisms enables the model to focus on the most clinically relevant areas, ultimately improving segmentation accuracy and reliability. This development not only aims to address the current gaps in medical image segmentation but also seeks to facilitate the integration of AI-driven

* Corresponding author.
E-mail addresses: engineerahmed204@gmail.com (A. Alweshah), r.barzamini.eng@iauctb.ac.ir (R. Barzamini), fhajati@une.edu.au (F. Hajati), Shoorangiz@iau.ac.ir (S.S.S. Farahani), m.arabian.eng@iauctb.ac.ir (M. Arabian), b.sohani@lboro.ac.uk (B. Sohani).

https://doi.org/10.1016/j.fraope.2024.100182
Received 26 June 2024; Received in revised form 11 October 2024; Accepted 11 November 2024
Available online 17 November 2024
2773-1863/Crown Copyright © 2024 Published by Elsevier Inc. on behalf of The Franklin Institute. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

tools into clinical workflows, thereby enhancing diagnostic accuracy and patient outcomes.

Medical image segmentation is one of the most important and challenging tasks in biomedical image analysis, as it can affect the diagnosis and treatment of various diseases and conditions [3–6]. According to a recent report, the global medical image analysis software market is expected to reach $4.51 billion by 2025, with a compound annual growth rate of 8.1 % [1]. However, medical image segmentation faces several difficulties, such as the variability of anatomical structures, the noise and artifacts in medical images, and the need for accurate and robust segmentation methods for diagnosis and treatment planning [2].

Semantic segmentation is a crucial task in medical image analysis that involves assigning a label to each pixel in an image to delineate anatomical structures or pathological regions. Deep learning techniques have become widely applied in the field of medical image segmentation [7,8]. In recent years, deep learning-based approaches [9–14] have shown significant promise in medical image segmentation, with the UNet architecture being a popular choice due to its ability to capture both local and global contextual information. The research on R-UNet is consistent with broader trends in medical image analysis, particularly in the development of deep learning methodologies to enhance diagnostic accuracy and streamline clinical workflow. However, the standard UNet architecture [15] does not explicitly model temporal dependencies, which can limit its ability to capture long-range contextual information in medical images. This is also a major issue in most UNet methods [16, 17].

To overcome this limitation, a modified UNet architecture is introduced, incorporating recurrent blocks at each stage of the encoder and decoder pathways. By incorporating recurrent layers and attention mechanisms, R-UNet addresses the demand for more context awareness and fine-grained segmentation, which is critical for tasks such as tumor identification and disease classification in medical images. Its ability to capture spatial-temporal dependencies and contextual information across images is pivotal in dealing with the inherent complexities of medical images, ultimately leading to more accurate segmentation results. This is in line with the overall trend towards automation in medical image analysis, with the aim of improving diagnostic accuracy while reducing the burden on healthcare professionals [18–20].

Furthermore, R-UNet's integration of attention mechanisms allows it to selectively focus on the most relevant regions of an image, which is crucial in the medical domain where subtle differences in tissue structure or texture can be indicative of disease. Attention mechanisms ensure that the model emphasizes these critical regions, potentially reducing false negatives in sensitive applications such as tumor detection. This targeted focus, combined with the recurrent layers' ability to capture long-range dependencies, enhances the network's capacity to model spatial relationships across multiple frames or slices of medical data, such as MRI or CT scans. The result is a more coherent and contextually aware segmentation, improving the model's performance in segmenting irregular and complex anatomical structures.

In the context of medical image analysis, the need for robust, automated tools like R-UNet is growing due to the sheer volume of medical images generated daily. Manual analysis by radiologists is time-consuming, prone to human error, and subject to variability between professionals. By leveraging deep learning models like R-UNet, healthcare systems can streamline the diagnostic process, offering consistent, reproducible results while maintaining high accuracy. This shift towards automation has the potential to enhance diagnostic workflows, allowing clinicians to focus on more complex cases or treatment decisions, ultimately improving patient outcomes. The ongoing research and development in this field are driving a new era of precision medicine, where AI-assisted tools complement human expertise to deliver faster, more reliable diagnoses.

In addition, the proposed approach involves additional skip connections that improve the extraction of features by emphasizing the importance of high-level features. To evaluate the performance of the proposed R-UNet architecture, different experiments are conducted on several medical image segmentation datasets, including the liver dataset, brain tumor dataset, and Mitochondria dataset. The experimental results demonstrate that the R-UNet architecture outperforms the standard UNet architecture and several other state-of-the-art semantic segmentation models in terms of segmentation accuracy improvement by 1.08 %, 3.02 % 4.1 %, and 7.16 % for the liver, brain Tumor, COVID-19, and Mitochondria datasets, respectively [21,22].

Beyond accuracy improvements, the proposed R-UNet architecture also demonstrates its ability to generalize effectively across diverse medical imaging modalities. Each dataset used in the experiments presents unique challenges—such as organ-specific tissue variability in liver images, complex structural irregularities in brain tumors, and intricate patterns in Mitochondria microscopy images. The ability of R-UNet to consistently outperform other models across these varied datasets speaks to its versatility. In medical image segmentation, where datasets can differ significantly in terms of imaging techniques (e.g., MRI, CT, or microscopy), a model's ability to handle this variability is crucial for its adoption in real-world clinical settings. The enhanced skip connections and attention mechanisms of R-UNet enable it to adapt to these different imaging characteristics, which suggests that this architecture is not limited to a narrow range of applications but could be extended to a broad array of medical domains [15,21].

In addition to its practical benefits, the R-UNet's advancements also offer important theoretical contributions to the field of deep learning-based medical image segmentation. The combination of recurrent layers, attention mechanisms, and advanced skip connections represents a significant evolution in model design, addressing common challenges such as class imbalance, small object detection, and the segmentation of irregularly shaped structures. The results of the experiments indicate that these innovations can lead to tangible improvements in segmentation performance without excessively increasing the computational complexity of the model. This is particularly important in medical applications where real-time or near-real-time processing is often required. The balance between accuracy and efficiency makes R-UNet a promising candidate for integration into clinical workflows, potentially serving as a core component of automated diagnostic systems that aim to enhance patient outcomes by providing faster and more accurate analysis [17, 22].

Unlike previous approaches, the aim is to incorporate recurrent networks and attention mechanisms to capture temporal dependencies [23], particularly in medical imaging where anatomical structures may exhibit changes over time due to physiological processes or diseases [24]. Through recurrent U-Net architecture, time series data is accurately segmented, ensuring consistency and coherence across successive frames or slices. These dependencies are crucial for tasks like lesion tracking, where smooth and continuous segmentation is necessary [25]. By leveraging contextual information from multiple previous frames, recurrent connections enhance the model's understanding of spatio-temporal context, leading to improved segmentation accuracy, especially for evolving features.

In contrast to prior studies [26–28] that integrate recurrent blocks into the UNet framework for medical image segmentation, the newly proposed R-UNet architecture has several advantages.

- First; the proposed architecture uses recurrent blocks at each up-and-down step, which allows the model to capture temporal dependencies at different scales.
- Second; the proposed architecture incorporates supplementary skip connections that enhance feature extraction by prioritizing high-level features.
- Third; the integration of attention mechanisms in the proposed R-UNet further distinguishes it from previous architectures. While recurrent blocks allow the model to capture temporal dependencies across different layers, attention mechanisms improve the model's focus on the most critical regions of the image. This dual approach,

leveraging recurrent blocks for temporal dependencies and attention mechanisms for spatial prioritization, enables the model to better handle complex, multi-scale patterns often found in medical images. This is particularly important for tasks like tumor segmentation, where both local details and the overall context must be accurately captured to distinguish between malignant and benign regions. Prior studies, although effective, did not fully exploit the potential of combining attention with recurrent structures, which gives the proposed architecture an edge in terms of both precision and contextual awareness.

- Finally, the architecture is designed to be more computationally efficient compared to other complex segmentation models. By using recurrent blocks strategically at each step of the UNet, the R-UNet manages to capture long-range dependencies without introducing excessive computational overhead. This efficiency is crucial in medical image processing, where high-dimensional data, such as 3D MRI or CT scans, are common. The ability to handle such data efficiently while maintaining high segmentation accuracy is a significant improvement over prior methods, which often struggle with either scalability or processing speed. This balance of performance and efficiency makes the R-UNet architecture a strong candidate for real-world clinical applications, where rapid, reliable segmentation is essential for diagnostic decision-making.

The rest of the paper is organized as follows: In Section 2, the current existing work is explored on medical image segmentation with a focus on methods that implement U-NET-based architectures. The proposed method is presented in Section 3, followed by the performance evaluations in Section 4 and the conclusions in Section 5.

2. Related work

Semantic segmentation has been an active research area in computer vision, with numerous deep learning-based approaches proposed to tackle this task [29]. Early efforts to address this issue were made by Arena et al. [9], who demonstrated how image processing techniques, particularly using cellular neural networks, could enhance the accuracy of medical diagnoses. Their work focused on improving tumour and cancer detection through Microarray analysis, aiding in the identification of inherited mutations that predispose family members to certain types of cancer. Then, the well-known U-Net architecture was introduced by Ranneberger et al. [16,30] in 2015, and has become a popular choice for semantic segmentation due to its ability to capture both local and global contextual information. The UNet architecture consists of a contracting path that captures context and a symmetric expanding path that enables precise localization. However, the standard UNet architecture does not explicitly model temporal dependencies, which can limit its ability to capture long-range contextual information in complex scenes or in medical images [31,32].

To address this limitation, researchers have proposed several modifications to the UNet architecture [33,34]. More specifically, in medical image analysis, various feature ex-traction models have been utilized to extract relevant features, including U-Net, U-Net++, and U-Net3+. Also, several other modifications have been proposed to enhance the original U-Net architecture, such as incorporating attention modules, drop-out models, and pyramid architecture, all of which have demonstrated efficient seg-mentation performance for various medical applications, including liver segmentation, brain tumour segmentation, cancer cell segmentation, and COVID-19 segmentation [35].

The U-Net architecture functions as an auto-encoder model, extracting high-level details from the decoder layers through convolution layers and ReLU activation function, and low-level details from the corresponding encoder layers. U-Net++ improves semantic segmentation by incorporating nested networks of skip connections, strengthening the connection between the input image and the ground truth [36,37]. U-Net3+ detects fine-grained features and global semantic representation on a full scale by using smaller and same-scale features from the shallow layers [38,39]. These improvements aim to develop a robust feature extraction model to optimize semantic segmentation of medical images. Despite the successful performance of these approaches, the latest U-Net versions have significantly increased in complexity.

Other researchers have also proposed modifications to the UNet architecture that incorporate recurrent blocks to capture temporal features. For example, in Dou et al. (2018) [35,40] the Recurrent Residual Convolutional Neural Network (R2U-Net) architecture is proposed which uses recurrent blocks to capture long-range dependencies and residual connections to enhance feature reuse. Another modification involves the integration of recurrent blocks to capture dependencies over time. For instance, in Zhou et al. (2019) [27] the Recurrent UNet (RU-Net) architecture is proposed which uses bidirectional LSTM layers to propagate features across time and capture long-range contextual information. More recently, researchers have proposed other variations of the UNet architecture that incorporate recurrent blocks to capture temporal dependencies. For example, in [29] the Recurrent Convolutional UNet (RCU-Net) architecture is proposed which uses convolutional LSTM layers to capture both spatial and temporal dependencies. Overall, these recurrent methods have shown potential improvement in the field of medical image segmentation.

In [41], a hybrid forecasting model is presented based on adaptive neuro-fuzzy inference system (ANFIS) and mutation-based Bees Algorithm (mBA)to determine the number of confirmed cases for the upcoming 10 days based on the earlier confirmed cases found in India. The meta-heuristic Bees Algorithm (BA) is modified using 4 types of mutation, and mBA is used to enhance the performance of ANFIS by optimizing its parameters.

Most recent advanced studies including the method proposed in [35] reduce network parameters while enhancing feature extraction capabilities through the use of channel reduction and a convolutional block attention module (CBAM). UniverSeg [36] employs few-shot learning techniques to tackle unseen medical segmentation tasks, applying a CrossBlock mechanism to transfer information from the example set to new images. Similarly, in [23] and [38] a genetic algorithm is used to automatically design a U-Net variant, incorporating residual connections and a cascade structure, respectively.

The proposed R-UNet architecture stands out from similar models due to its integration of recurrent convolutional layers, which enhance the network's ability to capture spatial-temporal dependencies and contextual information across the entire image. This incorporation of recurrent layers allows the network to maintain temporal memory and improve gradient flow during back-propagation. By combining the strengths of the attention mechanism for fine-grained detail extraction with the recurrent layers' capability to understand spatial hierarchies and dependencies, R-UNet achieves superior segmentation performance. This approach not only enhances feature propagation and contextual understanding but also makes the architecture highly adaptable to various tasks, including those involving sequential data like video segmentation. Empirical results demonstrate that R-UNet often surpasses traditional U-Net and similar architectures in accuracy and robustness, making it particularly effective for complex image segmentation tasks requiring precise boundary delineation and comprehensive context awareness.

3. The proposed method

In this section, the proposed Recurrent UNet (R-UNet) architecture is described for the semantic segmentation of medical images. R-UNet is an extension of the popular U-Net architecture that incorporates recurrent blocks to capture temporal dependencies. R-UNet combines convolutional and recurrent blocks in a complementary manner to capture both spatial and temporal dependencies in medical images.

Fig. 1 illustrates the architecture of the proposed method. It

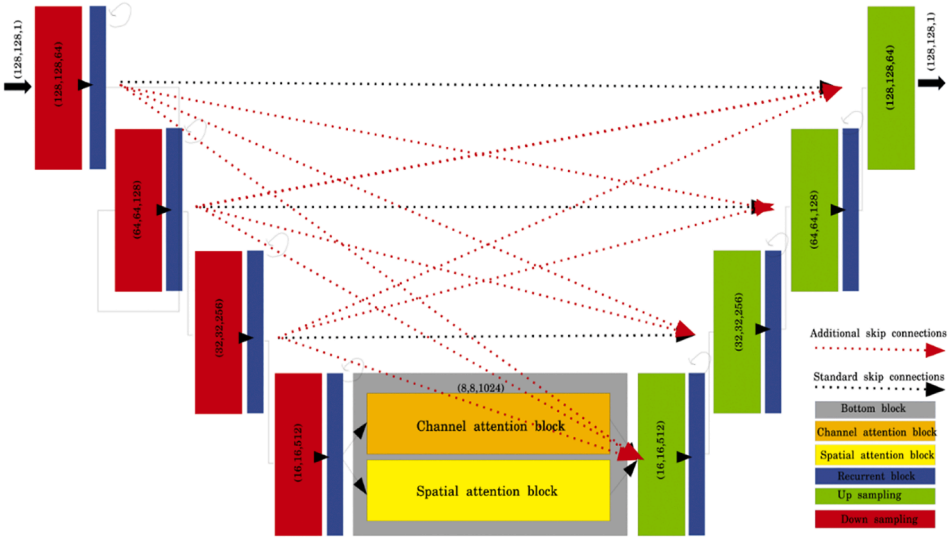


Fig. 1. The proposed Recurrent UNet (R-UNet) architecture.

comprises up-and-down convolutional blocks, recurrent blocks, attention blocks, and complementary skip connections. This design effectively captures both spatial and temporal dependencies in medical images. Algorithm 1 provides a detailed description of the proposed model. The process begins with input image dimensions (HEIGHT, WIDTH, CHANNELS) and outputs a 2D binary mask. Initially, the inputs are processed through a series of convolutional and recurrent layers, which increase in depth while the spatial dimensions decrease, accompanied by max-pooling layers. At the deepest layer, channel and spatial attention blocks are applied to enhance model performance. Subsequently, the model employs up-sampling layers that concatenate feature maps from corresponding down-sampling layers, applying convolutional and recurrent blocks to improve spatial resolution. The final output layer utilizes a sigmoid activation function to produce a binary segmentation mask.

3.1. Architecture

The R-UNet architecture consists of an encoder, a recurrent block, and a decoder. Fig. 1 illustrates the architecture of the proposed method in more detail.

The encoder part of the architecture comprises a series of

Algorithm 1
R-UNet model with attention channels, dropout, and L2 regularization.

Input: Input image dimensions (HEIGHT, WIDTH, CHANNELS)
Output: R-UNet model
// Input layer
1 inputs ← Input(HEIGHT, WIDTH, CHANNELS);
// Downsampling layers
2 down1 ← recurrent_block(inputs, 64, 3, dropout=0.5, kernel_regularizer=12(0.01));
downp1← MaxPooling2D()(down1);
3 down2 ← recurrent block(downp1, 128, 3, dropout=0.5,
Kernel_regularizer=12(0.01)); downp2 ← MaxPooling2D()(down2);
4 down3 ← recurrent_block(downp2, 256, 3, dropout=0.5,
kernel_regularizer=12(0.01)); downp3 ← MaxPooling2D()(down3);
5 down4 ← recurrent_block(downp3, 512, 3, dropout=0.5,
kernel_regularizer=12(0.01)); downp4 ← MaxPooling2D()(down4);
// Bottom layer
6 bottom ← recurrent_block(downp4, 1024, 3, dropout=0.5,
Kernel_regularizer=12(0.01)); bottom ← channel_attention(bottom); bottom ←
spatial_attention(bottom);
// Upsampling layers
7 up1 ← Conv2DTranspose(512, 2, strides=(2, 2)) (bottom); up1 ← concatenate([up1, down4]);
up1 ← recurrent_block(up1, 512, 3, dropout=0.5, kernel_regularizer=12(0.01);
8 sup2 ← Conv2DTranspose(256, 2, strides=(2, 2)) (up1); up2← concatenate([up2,down3]);
up2 ← recurrent_block(up2, 256, 3, dropout=0.5,
kernel_regularizer=12(0.01));
9 up3 ← Conv2DTranspose(128, 2, strides=(2, 2)) (up2); up3 ← concatenate([up3, down2]);
up3 ← recurrent_block(up3, 128, 3, dropout=0.5,
kernel_regularizer=12(0.01));
10 up4 ← Conv2DTranspose(64, 2, strides=(2, 2)) (up3); up4 ← concatenate([up4, down1]);
up4 ← recurrent_block(up4, 64, 3, dropout=0.5,
kernel_regularizer=12(0.01));
// Output layer
11 outputs ← Conv2D(1, (1, 1), activation='sigmoid')(up4);
// Model definition
12 model ← Model(inputs=inputs, outputs=outputs);
13 **return** model;

convolutional layers followed by max-pooling layers. These layers progressively reduce the spatial dimensions while increasing the depth of the feature maps. The purpose of the encoder is to extract hierarchical features from the input images [27,42].

Following the encoder, the recurrent block is introduced to capture temporal dependencies in the feature maps. This block incorporates recurrent connections that allow information to persist across different time steps, enabling the model to capture long-range dependencies [43]. The recurrent block is strategically placed within the network to enhance its ability to capture temporal information while preserving spatial details [18,44].

Finally, the decoder part of the architecture consists of up-sampling layers combined with skip connections from the encoder. These skip connections help preserve spatial information and recover finer details during the up-sampling process. The de-coder gradually increases the spatial dimensions while decreasing the depth of the feature maps, ultimately producing the segmentation mask [18–20].

By integrating convolutional and recurrent blocks in a unified architecture, R-UNet effectively addresses the challenges of capturing both spatial and temporal dependencies in medical images, leading to improved segmentation performance [45,46].

Additional skip connections (red colors in Fig. 1) are added to boost feature robustness and improve the detection of fine-grained details. Adding more skip connections facilitates easier gradient propagation, reducing gradient vanishing and improving training stability, especially in deeper networks. While this increases model complexity, it enhances the combination of multi-scale features, leading to improved feature representation and segmentation accuracy.

3.2. Up and down blocks

The encoder extracts features from the input image using a series of convolutional blocks. Each block consists of double convolution blocks and each block consists of several layers as follows:

1. 2D convolution layer with same size padding.
2. Batch normalization layer that normalizes the output of the previous layer by subtracting the batch mean and dividing by the batch standard deviation.
3. Rectified Linear Unit (Relu) activation layer to avoid the vanishing gradient problem.
4. Max-pooling or Up Sampling layers to decrease or increase the parameter size, respectively.

3.3. Recurrent blocks

The recurrent block is a key component in the proposed Recurrent UNet (R-UNet) architecture. By incorporating recurrent blocks into the U-Net architecture, the model gains the capacity to process temporal information within static images, leveraging contextual understanding. With memory cells retaining information from previous steps, the model can recall and combine features across different parts of the image, enhancing coherence in understanding overall structure and relationships. Additionally, iterative processing within static images offers a smoothing effect, enhancing consistency in segmentation maps and reducing noise and false predictions, especially in scenarios where changes over time are critical for clinical interpretation and decision-making. Each recurrent block is composed of a convolutional layer, followed by a gated recurrent unit (GRU) layer. The convolutional layer serves as a feature extractor, while the GRU layer captures temporal dependencies between adjacent feature maps. The GRU layer contains two gates: the reset gate and the update gate, which are responsible for resetting the cell state and updating the output state, respectively. While GRU layers introduce computational overhead, they significantly enhance the model's ability to capture spatial and temporal dependencies, leading to improved segmentation accuracy. The key is to balance the trade-off between accuracy and efficiency. To avoid overfitting, drop-down layers and L2 regularization layers are added as described in Algorithm 1.

Let h_t denotes the feature map at the time step t, x_t denote the input to the recurrent block at the time step t, and z_t denote the output of the recurrent block at the time step t. The recurrent block can be formulated as follows:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$\boldsymbol{r_t} = \boldsymbol{\sigma}(\boldsymbol{W_r} * \boldsymbol{x_t} + \boldsymbol{U_r} * \boldsymbol{h_{t-1}} + \boldsymbol{b_r}), \ (\textbf{\textit{Reset gate}}) \tag{2}$$

$$\boldsymbol{u_t} = \boldsymbol{\sigma}(\boldsymbol{W_u} * \boldsymbol{x_t} + \boldsymbol{U_u} * \boldsymbol{h_{t-1}} + \boldsymbol{b_u}), (\textbf{\textit{Update gate}}) \tag{3}$$

$$\boldsymbol{c_t} = \boldsymbol{tanh}(\boldsymbol{W_c} * \boldsymbol{x_t} + \boldsymbol{r_t} * \boldsymbol{U_c} * \boldsymbol{h_{t-1}} + \boldsymbol{b_c}), (\textbf{\textit{New cell state}}) \tag{4}$$

$$\boldsymbol{h_t} = (1 - \boldsymbol{u_t}) * \boldsymbol{h_{t-1}} + \boldsymbol{u_t} * \boldsymbol{c_t}, (\textbf{\textit{Output state}}) \tag{5}$$

where σ is the sigmoid activation function, * denotes the matrix multiplication, and $\boldsymbol{b}$ denotes the bias term. $\boldsymbol{W}$ and $\boldsymbol{U}$ denote the weight matrices, which are learnable parameters of the recurrent block. The reset gate $\boldsymbol{r_t}$ determines how much of the previous hidden state $\boldsymbol{h_{t-1}}$ to forget, while the update gate u t determines how much of the new cell state $\boldsymbol{c_t}$ to retain in the output state $\boldsymbol{h_t}$. The new cell state $\boldsymbol{c_t}$ is computed by applying a ***tanh*** activation function to the sum of the input $\boldsymbol{x_t}$ and te reset gate $\boldsymbol{r_t}$ multiplied by the previous hidden state $\boldsymbol{h_{t-1}}$. The output state $\boldsymbol{h_t}$ is a weighted combination of the previous hidden state $\boldsymbol{h_{t-1}}$ and the new cell state $\boldsymbol{c_t}$, controlled by the update gate $\boldsymbol{u_t}$.

3.3.1. Bottom block

The bottom block in the proposed Recurrent UNet architecture consists of an attention mechanism. The output of the recurrent block from the previous step is then passed through a channel attention layer and a spatial attention layer. The channel attention layer captures inter-channel dependencies by computing channel-wise feature responses and weighting them based on their importance [44]. The spatial attention layer captures inter-spatial dependencies by computing a spatial mask that highlights the most informative regions of the feature maps. These attention mechanisms help to selectively emphasize relevant features and suppress irrelevant ones, which improves the quality of the segmentation results [41].

Among the various attention models available, the Convolutional Block Attention Module (CBAM) stands out as a lightweight, feedforward convolutional neural network attention module that can be seamlessly integrated into any CNN architecture for end-to-end training. The CBAM structure consists of two main components: channel attention and spatial attention. The combination of channel and spatial attention in CBAM enhances the extraction of expressive feature information, which improves the contextual information retrieval from images of different scales.

In the proposed model, the bottom layer receives a feature map $\boldsymbol{h_t} \in \mathbb{R}^{C \times H \times W}$, which is processed through a convolution operation and then fed into the CBAM. First, the channel attention map $\boldsymbol{A_C} \in \mathbb{R}^{C \times 1 \times 1}$ is computed along with the channel-refined feature map $\boldsymbol{h_A}$. Next, the spatial attention map $\boldsymbol{A_S} \in \mathbb{R}^{1 \times H \times W}$ is calculated, resulting in the final feature map $\boldsymbol{h_S}$. The corresponding formulas for these calculations are as follows:

$$h_A = A_C\ (h_t) \otimes\ h_t \tag{6}$$

$$h_S = A_S\ (h_t) \otimes\ h_t \tag{7}$$

where $\otimes$ represents the element-wise multiplication corresponding to the feature matrix.

4. Performance evaluation

In this section, the performance of the proposed model is evaluated using several datasets, model training parameters, evaluation matrix and the final performance results compared to other methods. The model is trained using 50 batch size and 250 epochs on the NVIDIA GeForce RTX 3090 with a 24 GB memory machine.

4.1. Datasets

Six different medical imaging datasets were chosen to demonstrate the effectiveness of the proposed method. Examples of these datasets can be seen in Fig. 2.

1. The liver dataset [47] was used in a competition to identify the liver region by the IEEE International Symposium on Biomedical Imaging in 2017. The dataset consists of 131 3D CT scans, where 120 images are used for training and 11 for testing.
2. The brain tumor dataset [48] provides sub-regions from the Brain Tumor segmentation benchmark challenge in 2018. This dataset consists of 210 high-grade glioma (HGG) images and 75 low-grade glioma (LGG) images.
3. The Mitochondria image dataset [49] was also selected to measure the performance of the proposed method. This dataset has 165 images for each category Test, Test-masks, Train and Train-mask.
4. The lung CT COVID-19 images dataset was collected for the purpose of model evaluation. This dataset consists of 72 images, where 50 images are used for training and 22 images for testing. The images are pre-processed (0-255 pixel values) and resized (224×224 pixels). The input to the model during training was 3-channel images and the output was 1-channel binary images containing the COVID-19-affected region and background.
5. Nuclei Instance Segmentation in H&E-Stained Histological Images (NuInsSeg). The NuInsSeg dataset contains more than 30k manually segmented nuclei from 31 human and mouse organs and 665 image patches extracted from H&E-stained whole slide images. It also provides ambiguous area masks for the entire dataset to show in which areas manual semantic/instance segmentation were impossible.
6. The lung Segmentation dataset [50] is a subset of the Lung Nodule Analysis (LUNA) and Data Science Bowl competition in 2017.

4.2. Model parameters

1. Loss function; the sigmoid loss function is employed for several reasons. Firstly, it facilitates the estimation of prediction probabilities, which is crucial for binary segmentation tasks like ours. Secondly, the sigmoid loss function offers smoother gradients compared to other loss functions, promoting more stable and faster convergence during training. This smoother con-vergence trajectory often translates to better optimization performance and ultimately, improved model accuracy. Although cross-entropy loss is generally more robust and flexible, particularly for handling class imbalances and multi-class problems, the sigmoid loss function is selected for its faster and more effective convergence in our binary segmentation task involving two classes.
2. Adam optimizer; The choice of optimizer plays a pivotal role in the training of deep neural networks. In this study, Adam optimizer is used due to its robustness in handling noisy gradients, a common occurrence in deep learning training, especially in complex architectures like the proposed work. Adam's adaptive learning rate mechanism enables it to adjust learning rates for each parameter individually, leading to faster convergence and more stable training dynamics. Moreover, Adam optimizer tends to exhibit better generalization capabilities, effectively navigating the high-dimensional parameter space to find optimal solutions.
3. Learning rate and weight decay; hyperparameters are meticulously tuned to ensure optimal training performance. The learning rate, set to $1 \times 10^{-4}(0.0001)$, strikes a balance between rapid convergence and avoiding overshooting the minima. This value is chosen empirically through experimentation and has proven effective in achieving desirable training outcomes. Additionally, weight decay regularization is applied with a coefficient of $2 \times 10^{-3}(0.002)$. This regularization term aids in preventing overfitting by penalizing large weight values, thereby encouraging the model to learn simpler and more generalizable representations [51].

These choices collectively contribute to the efficacy of the training regimen, fostering the development of a robust and accurate segmentation model across various medical imaging datasets.

4.3. Evaluation metric

Various metrics are presented which are commonly used to evaluate and report the results of segmentation. These metrics include Intersection over Union (IoU), The Panoptic Quality (PQ), the Aggregate Jaccard Index (AJI), and the DICE similarity coefficient/score (also known as the DICE coefficient, which is a statistical measure used to gauge the similarity between two sets of data). These metrics are defined using terms such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) which are calculated based on the number of predicted pixels. The following evaluation metrics are calculated for testing purposes:

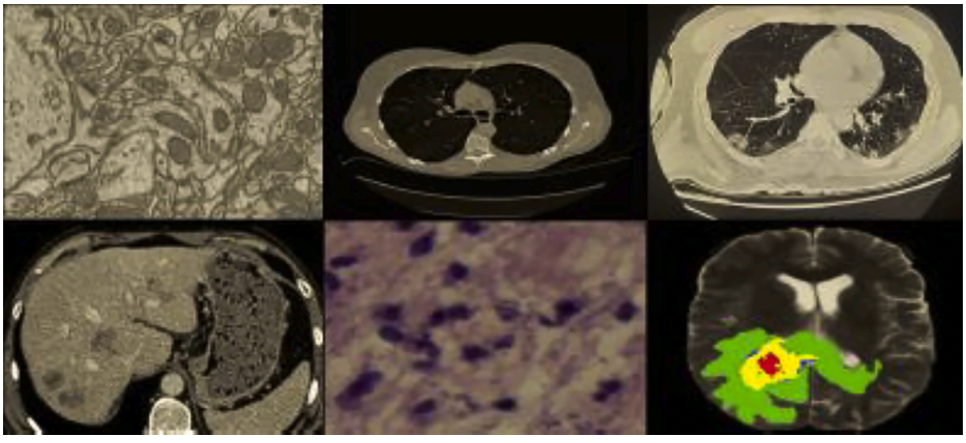


Fig. 2. Samples from different datasets used in the experiments.

$$Intersection\ over\ Union\ (IoU) = \frac{TP}{TP + FP + FN} \tag{8}$$

$$The\ Panoptic\ Quality\ (PQ) = \frac{IoU}{TP + \frac{1}{2}FP + \frac{1}{2}FN} \tag{9}$$

$$Aggregate\ Jaccard\ Index(AJI) = \frac{TP}{\sum_{i=1}^{N}|TP_i \cup FP_i \cup FN_i| + FP + FN} \tag{10}$$

$$DICE = \frac{2TP}{TP + 2FP + FN}. \tag{11}$$

The matrices intend to measure the overlapping performance between the ground truth and predicted output from the model. The IoU directly measures the overlap between the predicted segmentation and the ground truth. The AJI is a more complex metric that can be seen as a generalization of IoU for multi-instance segmentation. PQ also takes into account both the overlap and the correct detection of instances. Whereas, the Dice coefficient emphasizes the overlap between the predicted segmentation and the ground truth, giving more weight to the true positive rate. In general, AJI and PQ are more complex measurements compared to IoU and the DICE. DICE focuses on spatial overlap, making it an intuitive and direct measure of accuracy, crucial in medical imaging for clinical relevance.

DICE Eq. (11) is used as the primary evaluation metric, to measure the agreement between predicted segmentation and ground truth. Ranging from 0 to 1, with 1 indicating perfect agreement.

4.4. Results and discussion

In this section, the performance of the proposed model is compared with various relevant models and provide an in-depth analysis of the findings. The results of the experiments, including visual comparison, training and validation accuracy/loss plots are also presented and their implications are discussed. Table 1 presents a comparison of the most relevant studies across various datasets, focusing on the DICE score.

The results in Table 1 indicate that the proposed network leverages the extracted features to effectively segment medical images, thereby highlighting its superiority. The proposed method improves the segmentation performance compared to existing state-of-the-art studies such as U-Net 3+ and U-Net++ by an average of 2 %, 3 %, 7 %, and 1 % for the Liver, Tumor, Mitochondria and Lung datasets respectively. The highest accuracy was achieved when using the Mitochondria dataset with over 97 % accuracy resulting in a margin improvement of over 7 % compared to other relevant studies. This efficient performance on the Mitochondria dataset specifically can be attributed to its nature as a time sequence dataset. Recurrent layers, such as those used in the proposed model, are particularly adept at handling sequential data. These layers excel at capturing temporal dependencies and patterns, which significantly enhances the model's ability to learn from and predict sequential data effectively. Additionally, the model was trained on this dataset without shuffling the data. This approach preserves the inherent temporal structure of the dataset, further allowing the recurrent layers to leverage the continuity in the data, thereby boosting the overall performance. To provide further insight into the performance of the proposed model, Fig. 3a presents examples from the Mitochondria dataset. These examples showcase the effectiveness of the proposed model, particularly its ability to accurately detect each pixel using recurrent blocks. The visualization demonstrates the model's capability to identify intricate structures within medical images, highlighting its potential as a promising solution for segmentation tasks.

Table 1
The output results of the proposed method on different datasets compared with state-of-the-art methods.

Study	Liver detection	Tumor detection	COVID-19	Mitochondria
DeepLabV3 [26]	92.17	81.33	N/A	N/A
PSPNet [27]	92.42	84.867	N/A	N/A
Attention U-Net [29]	93.41	85.41	N/A	86.99
U-Net++ [11]	94.75	85.94	88.05	89.21
U-Net 3+ [12]	96.75	86.15	90.83	90.36
The proposed R-UNet	97.83	89.17	94.73	97.02

R-UNet outperforms other models in medical image segmentation across diverse datasets, accurately capturing detailed structures in organs like liver, brain tumors, lung and Covid regions. This is due to the large, well-defined objects in these images. This characteristic makes the segmentation task easier, allowing to achieve higher accuracy. Fig. 4 shows the segmentation accuracy for a different set of images. Although some datasets like the Liver dataset are easy for the segmentation task, the proposed model demonstrates effective performance on a complex and challenging dataset such as the Mitochondria dataset. Challenges such as dataset variability and class imbalance need addressing, necessitating tailored model training and robust preprocessing for reliable performance. This finding highlights the importance of combining recurring features and integrating spatial-temporal mechanisms. While recurrent layers capture long-term dependencies and contextual information, attention mechanisms focus on and localize important features. This collectively enhances the feature extraction capabilities of the Unet model for segmentation tasks.

Moreover, the proposed R-UNet architecture's performance extends beyond accuracy metrics; it also emphasizes the model's robustness in handling diverse and challenging segmentation tasks. In addition to achieving high DICE scores, the model demonstrates consistent performance across different datasets, highlighting its adaptability to various imaging conditions and data distributions. The results indicate that the model can effectively generalize across different types of medical images, including those with varying noise levels, resolution, and contrast. This adaptability is essential in real-world clinical applications, where medical images can often exhibit significant variability. As a result, the R-UNet not only provides high accuracy but also maintains reliability, making it suitable for deployment in diverse clinical settings.

The model's ability to tackle class imbalance and dataset variability is particularly noteworthy. In many medical imaging scenarios, certain classes, such as tumors or other pathologies, may be underrepresented in the training data. The proposed R-UNet architecture addresses these challenges through its enhanced feature extraction capabilities and the incorporation of recurrent and attention mechanisms. By effectively capturing critical features and contextual relationships, the model can focus on minority classes during training, leading to improved segmentation outcomes. This approach highlights the importance of considering class distribution when designing segmentation models and suggests that R-UNet could be fine-tuned further to enhance its performance on datasets with pronounced class imbalance.

Furthermore, the visual comparison of segmentation results reveals that R-UNet not only improves quantitative metrics but also enhances qualitative outcomes. Figures illustrating the segmentation masks demonstrate the model's ability to delineate intricate anatomical structures with remarkable precision. Such visual insights are crucial in medical image analysis, where understanding the model's decision-making process can help clinicians gain confidence in automated segmentation tools. The clear visualization of segmentation boundaries reinforces the potential for R-UNet to assist healthcare professionals in diagnostics, treatment planning, and monitoring disease progression. As medical imaging continues to evolve, integrating advanced deep learning techniques like those employed in R-UNet could significantly enhance clinical workflows and patient care.

4.5. Visual illustration

To compare the visualization with other methods, a visual

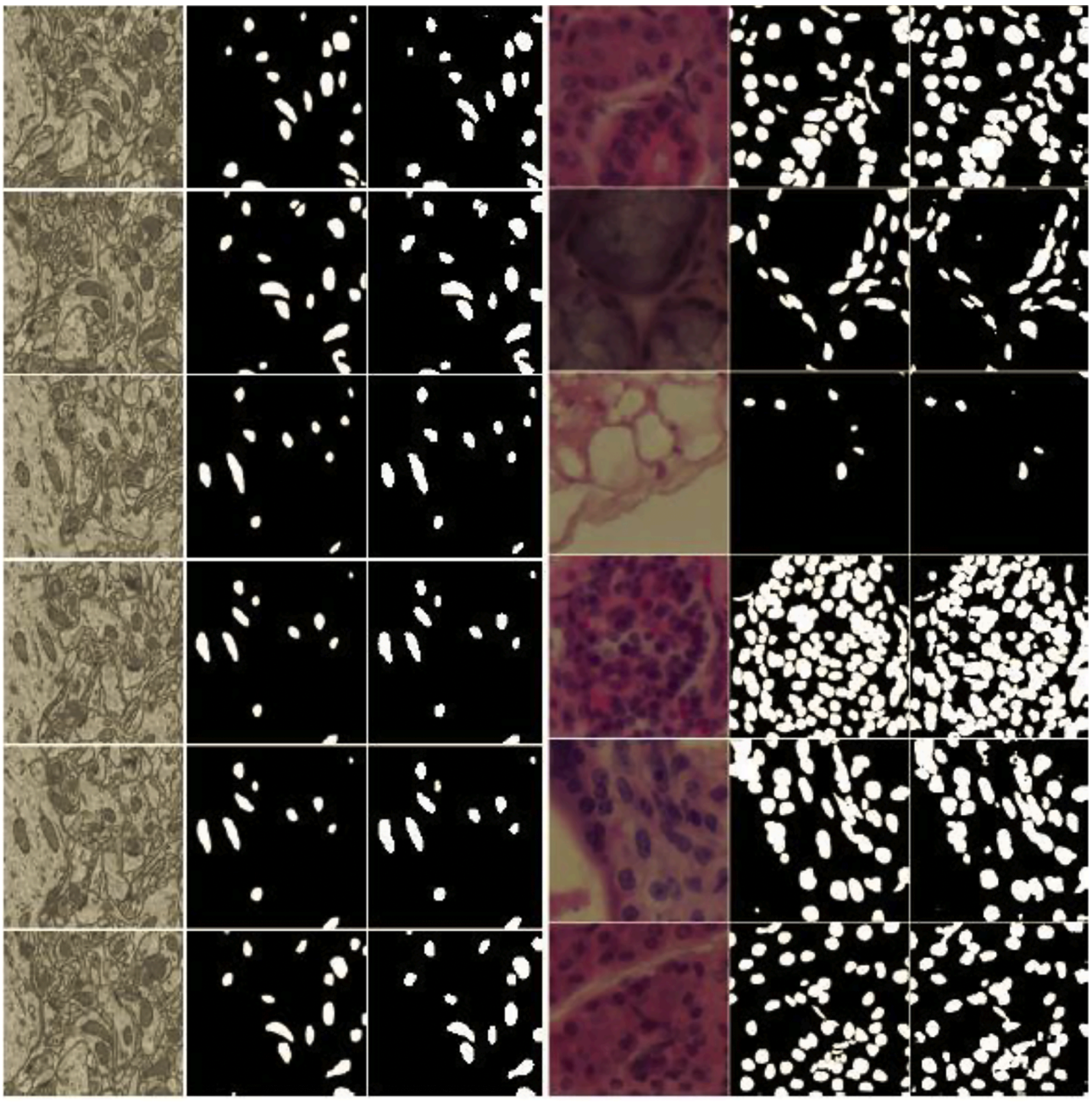


Fig. 3. Examples from the Mitochondria and NuInsSeg dataset the testing image (left column), ground truth (middle column) and R-UNet prediction (right column).

comparison is provided based on the lung dataset, as shown in Fig. 5. Various methods were applied to the lung dataset using a threshold of 0.5 to generate the final binary mask. While all methods achieved high segmentation accuracy, the proposed model produced the most detailed results by minimizing false positive predictions compared to other models.

In addition to the detailed results observed in Fig. 5, the qualitative assessment underscores the importance of visual interpretability in evaluating segmentation performance. The proposed R-UNet model not only excels in producing precise segmentation masks but also effectively preserves the intricate boundaries of the lung structures, which are often challenging due to overlapping textures and varying intensities in the images. This ability to maintain clarity around critical regions, such as the vascular structures and airways, is vital for applications in pulmonary diagnostics and treatment planning. By reducing false positives, R-UNet enhances clinical decision-making, allowing radiologists and healthcare providers to focus on relevant pathologies without being misled by extraneous information. Furthermore, this level of detail in segmentation results contributes to improved downstream tasks, such as volumetric analysis or automated lesion detection, thereby highlighting R-UNet's potential as a versatile tool in medical image analysis and its applicability to other challenging datasets beyond lung images.

4.6. Training monitoring

Throughout the training process, the accuracy and loss functions are closely monitored. Fig. 6a illustrates the accuracy curve depicting the model's performance on the validation set across each epoch. The curve illustrates a steady increase in accuracy as training progresses with stabilization observed around epoch 100. The training is extended

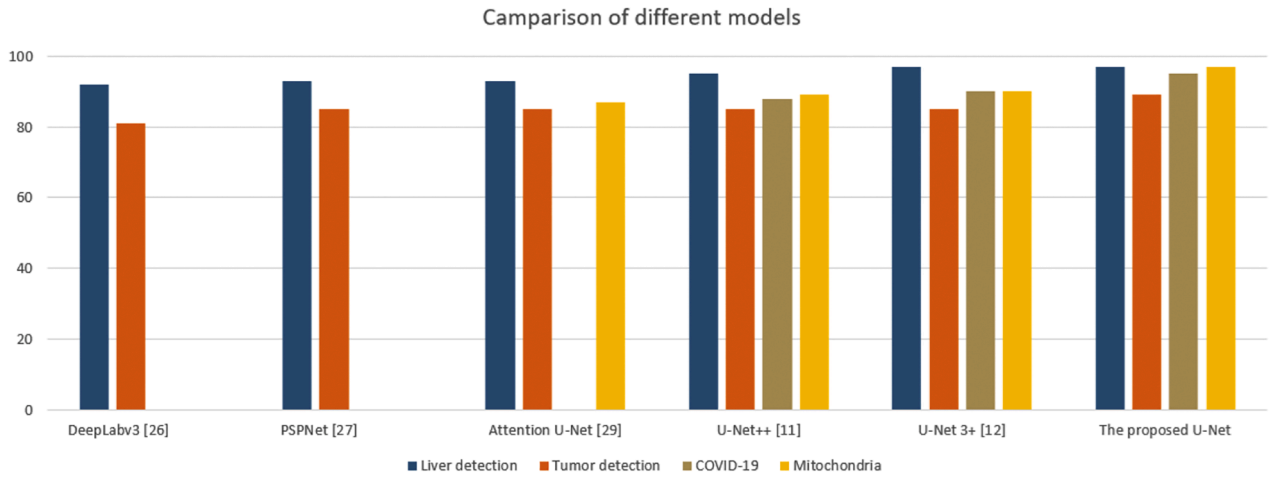


Fig. 4. Comparison of different dataset accuracies using our proposed model.

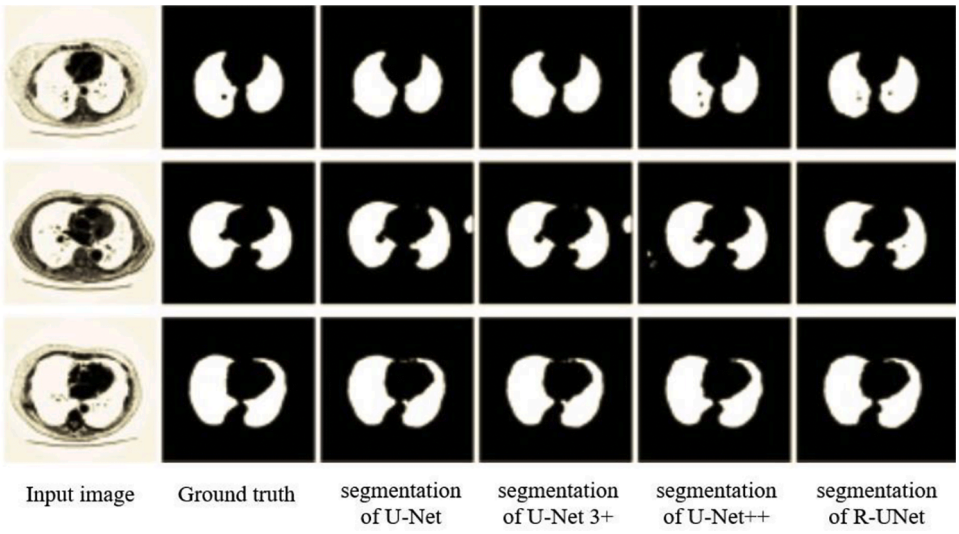
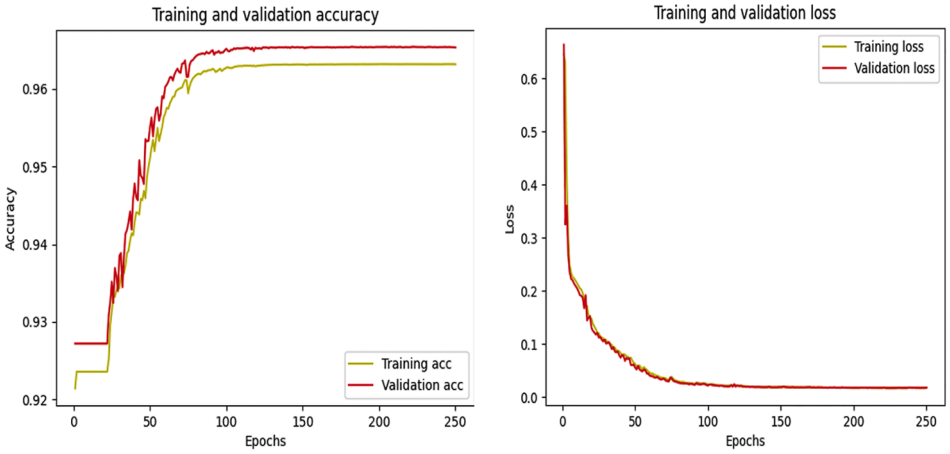


Fig. 5. Visual comparison of different methods using lung segmentation datasets.



(**a**) Accuracy (**b**) Loss

Fig. 6. Accuracy and loss plots during the training session of the Mitochondria dataset.

beyond the point of stability to explore the potential for further improvement or fine-tuning of the model. The exploration of the model's learning trajectory is prioritized to gain a comprehensive understanding of its behavior and performance characteristics. Fig. 6b displays the loss curve, indicating the model is able to minimize the difference` between the predicted and actual outputs during training. The decreasing trend in loss over epochs signifies the model's learning process and continual improvement in predictions.

4.7. Real-time implementation

Regarding real-time processing in clinical settings, R-UNet can achieve real-time processing without GPU acceleration, relying on fast CPUs and SSD storage for efficient operation. This ensures timely and accurate segmentation crucial for clinical decision-making, though performance may vary based on machine resources. Future improvements could involve enhancing the R-UNet architecture with transformer blocks and incorporating unsupervised techniques with Bayesian probability to reduce the need for manual data annotation.

Furthermore, the potential integration of advanced techniques, such as semi-supervised learning and self-supervised learning, could enhance R-UNet's capability to leverage unannotated data effectively. By utilizing a smaller set of labeled images alongside a larger pool of unlabeled data, these methods can improve the model's learning efficiency and performance, especially in scenarios where annotated datasets are scarce or expensive to obtain. This approach allows the model to learn meaningful representations from the unlabeled data, subsequently refining its segmentation accuracy on the labeled set. Additionally, incorporating techniques like data augmentation and synthetic data generation can enrich the training process by creating a more diverse dataset, further improving the model's robustness against variations in imaging conditions. Collectively, these strategies not only aim to reduce reliance on extensive manual annotation but also ensure that R-UNet remains adaptable and resilient in the face of evolving medical imaging challenges, positioning it as a valuable asset in the ongoing advancement of automated diagnostic tools in healthcare.

4.8. Challenges and future applications

The limitations and challenges of the proposed R-UNet model include increased computational complexity resulting from the integration of recurrent layers and attention mechanisms, which may lead to longer training times and increased memory requirements, especially when working with large datasets or limited hardware resources. In addition, the model's dependency on high-quality training data for optimal performance and the risk of overfitting associated with its complexity, necessitate the implementation of stringent regularization techniques as shown in Algorithm 1.

Finally, the proposed R-UNet architecture can be applied to various fields, including fake image generation, microscopy image analysis, and medical and satellite image segmentation. It can be fine-tuned and adapted to address specific challenges, advancing medical imaging, biomedical research, and other image segmentation applications. This work builds on the foundation laid by previous research, demonstrating continued innovation and improvement in the field of medical image processing using CNNs.

4.9. Ablation experiments

The performance of the proposed R-UNet is investigated by testing different components and functionalities to understand the contribution of the model as follows:

4.9.1. Multi-organ segmentation

The proposed model is trained on the NuInsSeg dataset in order to verify performance on a wide range of different organs in one go. NuInsSeg dataset provides several human organs such as cerebellum, cerebrum (brain), colon (rectum), epiglottis, jejunum, kidney, liver, lung, melanoma, muscle, esophagus, palatine tonsil, pancreas, peritoneum, placenta, salivary gland, spleen, stomach (cardia), stomach (pylorus), testis, tongue, umbilical cord, and urinary bladder. In addition, the dataset has also Mouse organs including the cerebellum, cerebrum, colon, epiglottis, lung, melanoma, muscle, peritoneum, stomach (cardia), stomach (pylorus), testis, umbilical cord, and urinary bladder. The proposed method is trained on all these organs as one go. The proposed model shows a significant improvement of 84.2 % compared to the existing works [52] with an accuracy of 79.4 %. Visual results of the proposed method are provided in Fig. 3b to show the segmentation of different organs. Furthermore, in the next section (Section 4.9.2), the performance of different models and different evaluation functions are compared on the NuInsSeg dataset.

Moreover, the findings from the NuInsSeg dataset highlight the model's potential for application in personalized medicine and multi-organ analysis. As healthcare increasingly moves towards tailored treatment plans, the ability to accurately segment multiple organs in a single workflow can streamline diagnostic processes and enhance the effectiveness of therapeutic interventions. For instance, by efficiently identifying and delineating various organs, the proposed model can assist in creating comprehensive reports that summarize the health status of multiple anatomical structures simultaneously. This capability is particularly advantageous in oncology, where understanding the relationship between tumors and surrounding organs is crucial for treatment planning. Additionally, as the proposed model demonstrates its proficiency in handling diverse organ types, it opens avenues for future research exploring cross-species applications, such as translating findings from mouse models to human studies, ultimately bridging the gap between preclinical and clinical settings. This adaptability not only showcases the robustness of the proposed architecture but also signifies its potential to contribute meaningfully to advancements in medical imaging and diagnostics.

4.9.2. Multi-organ segmentation

To assess the outcomes of the NuInsSeg instance segmentation tasks, the Dice score, the aggregate Jaccard Index (AJI), and the panoptic quality (PQ) score are applied. The Dice score measures overall semantic segmentation performance, while the AJI and PQ scores also gauge a model's proficiency in distinguishing adjacent objects, thereby quantifying instance segmentation performance. Higher values in the Dice score, AJI, and PQ score signify superior performance. PQ and AJI are computed based on Eq. (9) and Eq. (10) respectively. Table 2 shows the outcome of three different evaluation matrices comparing the proposed method with the recent relevant studies.

4.9.3. Multi-organ segmentation

The attention mechanism plays a crucial role in enhancing the performance of the proposed model by allowing it to focus on the most relevant parts of the input data. A series of experiments is conducted to assess the impact of the attention mechanism on model performance.

These experiments in Table 3 involve the following cases:

1. R-UNet without attention mechanism nor recurrent blocks.

Table 2
Comparison between different studies using different evaluation metrics (%) on NuInsSeg dataset.

No.	Study	DICE	AJI	PQ
1	Two-stage U-Net [24]	76.6	52.7	47.2
2	U-Net+Watershed [12]	78.8	50.5	42.8
3	dual decoder U-Net [52]	79.4	55.9	51.3
4	Attention U-Net [23]	80.5	45.7	36.4
5	Residual attention U-Net [18]	81.4	46.3	36.9
6	The proposed R-UNet	84.2	61.9	54.7

Table 3
Evaluation of the proposed method using different scenarios on different datasets.

Case.	Liver	Tumor	COVID-19	Mitochondria	Lung	NulnsSeg
1	92.3	82.9	88.7	90.6	92.7	79.8
2	95.9	86.7	93.3	95.8	96.9	83.8
3	94.4	84.1	90.73	92.2	93.8	80.5
4	97.83	89.17	94.73	97.2	97.22	84.2

2. R-UNet without an attention mechanism.
3. R-UNet without recurrent blocks.
4. R-UNet with attention mechanism and recurrent blocks.

The evaluation of the R-UNet model in Table 3 across various scenarios demonstrates the significant impact of both the attention mechanism and recurrent blocks on performance across multiple datasets. Case 1, which lacks both attention mechanisms and recurrent blocks, consistently shows the lowest performance metrics. Incorporating either the attention mechanism (Case 2) or recurrent blocks (Case 3) independently improve performance across all datasets. Specifically, Case 2, which includes the recurrent blocks but not the attention mechanism, achieves noticeable improvements. The most substantial improvement is observed in Case 4, which integrates both the attention mechanism and recurrent blocks. This combination achieves the highest accuracy across all datasets. These results highlight that the synergistic effect of combining attention mechanisms and recurrent blocks significantly enhances the R-UNet model's capability, suggesting that both components are crucial for maximizing performance in medical image segmentation tasks.

Furthermore, the ability of the proposed model to process a wide variety of organ types within a single framework also paves the way for the integration of artificial intelligence in routine clinical practice. By automating the segmentation process across multiple organs, healthcare professionals can save valuable time and resources, allowing them to focus more on interpreting results and patient care rather than manual segmentation tasks. This shift not only improves workflow efficiency but also minimizes human error associated with manual delineation, leading to more reliable and consistent results. Additionally, the insights gained from the model's segmentation outputs can be used to inform machine learning algorithms for predicting disease progression or response to therapy, thereby supporting a more proactive approach in patient management. As the landscape of medical imaging evolves with the incorporation of AI-driven tools, the proposed model stands at the forefront of this transformation, promising to enhance diagnostic accuracy and improve overall patient outcomes in a clinical setting.

5. Conclusion

Convolutional neural networks (CNNs) have emerged as the cornerstone for medical image analysis due to their unparalleled performance. Among the myriads of CNN, U-NET has consistently showcased remarkable results in medical image segmentation tasks. In this study, a novel architecture, (R-UNet) is introduced which integrates fully skip connections and recurrent blocks specifically tailored for medical image segmentation.

The proposed network, R-UNet, is an extension of the U-Net framework, strategically augmented with optimal connections to the decoder end and recurrent blocks at each up and down step. This design facilitates the capture of temporal dependencies and long-range contextual information crucial for accurate segmentation. Through rigorous evaluation, the efficacy of our method is demonstrated across several public datasets, including the Liver dataset, Brain Tumor dataset, Mitochondria dataset, Lung dataset, NuInsSeg dataset and introduce a new collection of COVID-19 images for evaluation purposes. The proposed method has been comprehensively validated and compared with other existing works. The experimental results show that our method outperforms the state-of-the-art methods by 1.08 %, 3.02 %, 4.1 %, 7.16 %, 1 % for the Liver, Brain Tumor, COVID-19, Mitochondria and Lung datasets, respectively. These findings underscore the potential of the proposed method to revolutionize medical image segmentation tasks and furnish medical professionals with a powerful tool for enhanced diagnostic accuracy and efficiency. The proposed R-UNet architecture applies to other fields such as fake image generation, microscopy analysis, and medical and satellite image segmentation. It can be fine-tuned for specific challenges, enhancing medical imaging and biomedical research.

CRediT authorship contribution statement

Ahmed Alweshah: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis. **Roohollah Barzamini:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Farshid Hajati:** Writing – review & editing, Resources, Formal analysis, Data curation, Conceptualization. **Shoorangiz Shams Shamsabad Farahani:** Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization. **Mohammad Arabian:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis. **Behnaz Sohani:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Funding acquisition, Data curation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, et al., The medical segmentation decathlon, Nature Commu-nications 13 (4128) (2022), https://doi.org/10.1038/s41467-022-26104-1.
[2] Li-Hao Jhan, Chia-Hong Hsieh, Chi-Yuan Wang, Segmentation of spine vertebral bodies in mri im-ages using T1, T2, and stir sequences by deep learning method, J. Med. Imag. Radiat. Sci. (2023).
[3] Q. Gao, Medical image segmentation algorithm based on deep learning and convolutional neural network, in: 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 2023, pp. 01–05.
[4] R. Gu, et al., CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imaging 40 (2) (Feb. 2021) 699–711.
[5] A.A. Nair, T.D. Tran, A. Reiter, M.A. Lediju Bell, One-step deep learning approach to ultrasound image formation and image segmentation with a fully convolutional neural network, in: 2019 IEEE International Ultrasonics Symposium (IUS), Glasgow, UK, 2019, pp. 1481–1484.
[6] S. Patel, An overview and application of deep convolutional neural networks for medical image segmentation, in: 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 722–728.
[7] Eshmam Rayed, S.M. Sajibul Islam, Sadia Islam Niha, Jamin Rahman Jim, Md Mohsin Kabir, M.F. Mridha, Deep learning for medical image segmentation: State-of-the-art advancements and challenges, Volume 47, 2024, 101504.
[8] Mahdiyeh Rahmani, Ali Kazemi, Maryam Jalili Aziz, Ebrahim Najafzadeh, Parastoo Farnia, Alireza Ahmadian, A systematic review on medical image segmentation using deep learning, Scientia Iranica (2024) in press.
[9] F. He, T. Liu, D. Tao, Why ResNet Works? Residuals generalize, IEEE Transactions on Neural Networks and Learning Systems 31 (12) (2020) 5349–5362.
[10] Md.Faysal Ahamed, Md.Munawar Hossain, Md. Nahiduzzaman, Md.Rabiul Islam, Md.Robiul Islam, Mominul Ahsan, Julfikar Haider, A review on brain tumor segmentation based on deep learning methods with federated learning techniques, Comput. Med. Imaging Graph. 110 (2023) 102313. December.
[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[12] Jongwat Cheewakul, N Yanaihap, T Yue-roh, N Kepan, W Thiangsook, Th Rojanaaumpawan, N Ina, A deep learning for automated segmentation of skeletal muscle at the third lumbar vertebra in computed tomography images, J. Med. Imag. Radiation Sci. (2022).
[13] Geoff Currie, K.Elizabeth Hawk, Eric Rohren, Alanna Vial, Ran Klein, Machine learning and deep learning in medical imaging: intelligent imaging, J. Med. Imag. Radiation Sci. (2019).

[14] Kuei-Fen Chen, Hsiao-Pei Sung, Shih-Yen Hsu, Tai- Been Chen, Chi-Yuan Wang, Deep learning-based classifi-cation of ultrasound images by different protocols, J. Med. Imag. Radiat. Sci. (2023).

[15] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image seg-mentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[16] K. Lan, J. Cheng, J. Jiang, X. Jiang, Q. Zhang, Modified UNet++ with atrous spatial pyramid pooling for blood cell image segmentation, Math. Biosci. Eng. MBE 20 (2023) 1420–1433.

[17] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.

[18] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.H. Han, Y.W. Chen, R. Tong, Mixed Transformer UNet for medical image segmentation, in: Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: Piscataway, SingaporeNJ, USA, 2022, pp. 2390–2394, 23–27 May 2022.

[19] Lei, T.; Sun, R.; Wan, Y.; Xia, Y.; Du, X.; Nandi, A.K. TEC-Net: Vision Transformer Embrace Convolutional Neural Networks for Medical Image Segmentation. arXiv 2023, arXiv:2306.04086.

[20] S. Sun, Z. Cao, H. Zhu, J Zhao, A survey of optimization methods from a machine learning perspective, IEEE Trans. Cybern. 50 (2019) 3668–3681.

[21] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, A.L Martel, Loss odyssey in medical image segmentation, Med. Image Anal. 71 (2021) 102035.

[22] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M.Swin-UNet Wang, UNet-like pure Transformer for medical image segmentation, in: Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, Springer, Berlin/Heidelberg, Germany, 2022, pp. 205–218, 23–27 October 2022.

[23] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Jian Wu, Unet 3+: A full- scale connected unet for medical image segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 1055–1059.

[24] M.Z. Alom, C. Yakopcic, T.M. Taha, V.K. Asari, Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net), in: NAECON 2018 - IEEE National Aerospace and Electronics Conference, Dayton, OH, USA, 2018, pp. 228–233.

[25] Zhongming Luo, Yu Zhang, Lei Zhou, Binge Zhang, Jianan Luo, Haibin Wu, Micro-vessel image segmentation based on the AD-UNet model, IEEE Access 7 (2019) 143402–143411.

[26] Jun Chen, Lianlian Wu, Jun Zhang, Liang Zhang, Dexin Gong, Yilin Zhao, Qiuxiang Chen, Shulan Huang, Ming Yang, Xiao Yang, Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography, Sci. Rep. 10 (1) (2020) 1–11.

[27] Wei Wang, Kaicheng Yu, Joachim Hugonot, Pascal Fua, Mathieu Salzmann, Recurrent U-Net for re-source-constrained segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2142–2151.

[28] X. Wu, L. Yang, Y. Ma, C. Wu, C. Guo, H. Yan, Z. Qiao, S. Yao, Y. Fan, An end-to-end multiple side-outputs fusion deep supervision network based remote sensing image change detection algorithm, Signal Process. 213 (2023) 109203.

[29] R. Ando, et al., Detection of Cell Nuclei using LadderNet, in: 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), Kitakyushu, Japan, 2020, pp. 467–470.

[30] S. Balachandran, X. Qin, C. Jiang, E.S. Blouri, A. Forouzandeh, M. Dehghan, D. Zonoobi, J. Kapur, J. Jaremko, K. Punithakumar, ACU2E-Net: A novel predict–refine attention network for segmentation of soft-tissue structures in ultrasound images, Comput. Biol. Med. 157 (2023) 106792.

[31] K. Lan, J. Cheng, J. Jiang, X. Jiang, Q. Zhang, Modified UNet++ with atrous spatial pyramid pooling for blood cell image segmentation, Math. Biosci. Eng. MBE 20 (2023) 1420–1433.

[32] A. Kar, M. Petit, Y. Refahi, G. Cerutti, C. Godin, J. Traas, Benchmarking of deep learning algorithms for 3D instance segmentation of confocal image datasets, PLOS Comput. Biol. 18 (2022) e1009879.

[33] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, C. Fan, SA-UNet: spatial attention U-Net for retinal vessel segmentation, in: 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 1236–1242.

[34] S.D. Deb, R.K. Jha, Modified Double U-Net Architecture for Medical Image Segmentation, IEEE Transactions on Radiation and Plasma Medical Sciences 7 (2) (2023) 151–162. Feb.

[35] Busra Emek Soylu, Mehmet Serdar Guzel, Gazi Erkan Bostanci, Fatih Ekinci, Tunc Asuroglu, Koray Acici, Deep-Learning-based approaches for semantic segmentation of natural scene images: a review, Electronics 12 (12) (2023) 2730, https://doi.org/10.3390/electronics12122730.

[36] Jongwat Cheewakul, et al., Improving semantic segmentation of medical images with lighter U-Net architectures, J. Med. Imag. Radiat. Sci. 26 (3) (2024) 431–440.

[37] J. Liu, X. Fan, J. Jiang, R. Liu, Z Luo, Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion, IEEE Trans. Circuits Syst. Video Technol. 32 (2021) 105–119.

[38] H. Huang, et al., UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 1055–1059.

[39] W. Liu, T. Tian, W. Xu, H. Yang, X. Pan, S. Yan, L. Wang, Phtrans: Parallelly aggregating global and local representa-tions for medical image segmentation, in: Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022), Singapore, 2022, pp. 235–244, 18–22 September.

[40] C. Li, Y. Tan, W. Chen, X. Luo, Y. He, Y. Gao, F. Li, ANU-Net: Attention-based nested U-Net to exploit full resolution features for medical image segmentation, Comput. Graph. 90 (2020) 11–20.

[41] S. Saif, P. Das, S Biswas, A hybrid model based on mBA-ANFIS for COVID-19 confirmed cases prediction and forecast, J. Inst. Eng. India Ser. B 102 (2021) 1123–1136.

[42] X. Jiang, C Xu, Deep learning and machine learning with grid search to predict later occurrence of breast Cancer metastasis using clinical data, J. Clin. Med. 11 (2022) 5772.

[43] M.E. Paoletti, X. Tao, L. Han, Z. Wu, S. Moreno-Álvarez, S.K. Roy, A. Plaza, J Haut, Parameter-free attention network for spectral-spatial hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 61 (2023) 5516817.

[44] Li, W.; Wang, X.; Xia, X.; Wu, J.; Xiao, X.; Zheng, M.; Wen, S. Sepvit: Separable vision transformer. arXiv 2022, arXiv:2203.15380.

[45] W. Liu, T. Tian, W. Xu, H. Yang, X. Pan, S. Yan, L. Wang, Phtrans: Parallelly aggregating global and local representa-tions for medical image segmentation, in: Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022), Singapore, 2022, pp. 235–244, 18–22 September.

[46] J.M. Valanarasu, P. Oza, I. Hacihaliloglu, V. Patel, Medical Transformer: Gated Axial-Attention for Medical Image Seg-mentation, in: Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI2021), Part I, Strasbourg, France 12901, 2021, pp. 36–46, 27 September–1 October.

[47] Leslie N. Smith, Cyclical Learning Rates for Training Neural Networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 464–472.

[48] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, The liver tumor segmentation benchmark (lits), Med. Image Anal. 84 (2023) 102680.

[49] Lucchi Aurelien, Li Yunpeng, Carlos Becker, Fua Pascal, Electron microscopy dataset, EPFL (2013).

[50] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[51] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, Pyramid scene parsing network, in: Pro-ceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[52] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Jian Wu, UNet 3+: A Full-Scale Connected UNet for Medical Image Segmen-tation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, https://doi.org/10.1109/ICASSP40776.2020.9053405.