International Journal of Network Dynamics and Intelligence Scilight

Article

UNet and Variants for Medical Image Segmentation

Walid Ehab, Lina Huang, and Yongmin Li*

Department of Computer Science, Brunel University London, Uxbridge, UB8 3PH, United Kingdom
* Correspondence: yongmin.li@brunel.ac.uk

Received: 22 September 2023
Accepted: 25 December 2023
Published: 26 June 2024

Abstract: Medical imaging plays a crucial role in modern healthcare by providing non-invasive visualisation of internal structures and abnormalities, enabling early disease detection, accurate diagnosis, and treatment planning. This study aims to explore the application of deep learning models, particularly focusing on the UNet architecture and its variants, in medical image segmentation. We seek to evaluate the performance of these models across various challenging medical image segmentation tasks, addressing issues such as image normalization, resizing, architecture choices, loss function design, and hyperparameter tuning. The findings reveal that the standard UNet, when extended with a deep network layer, is a proficient medical image segmentation model, while the Res-UNet and Attention Res-UNet architectures demonstrate smoother convergence and superior performance, particularly when handling fine image details. The study also addresses the challenge of high class imbalance through careful preprocessing and loss function definitions. We anticipate that the results of this study will provide useful insights for researchers seeking to apply these models to new medical imaging problems and offer guidance and best practices for their implementation.

Keywords: medical imaging; segmentation; performance analysis; UNet; Res-UNet; Attention Res-UNet

1. Introduction

Medical image segmentation is a critical aspect of medical image analysis and computer-aided diagnosis, which involves the partitioning of images into meaningful regions for identification of structures such as organs, tumors and vessels. Deep learning, with its ability to automatically extract complex features from vast medical image datasets, presents a promising solution to enhance segmentation accuracy. Note that challenges persist due to the diversity of medical domains, necessitating tailored approaches and evaluation metrics.

The primary goal of this paper is to comprehensively study the state-of-the-art deep learning methods with the focus on the UNet [1] and its variants (the Res-UNet [2] and attention Res-UNet [3]) which are renowned for their effectiveness in addressing complex medical image segmentation tasks.

The main objectives of this work are to apply the UNet model and its variants to solve a number of representative medical image segmentation problems by adapting different image pre-processing and model training techniques, identifying appropriate performance metrics, and evaluating the performance of these models. Hopefully, the findings of this study will offer useful guidance to researchers when applying these models to solve new medical imaging problems.

The remainder of this paper is orgainised as follows. The problems of medical imaging and previous studies on segmentation, particularly medical image segmentation, are reviewed in Section 2. The details of the UNet, its variants, and evaluation methods are discussed in Section 3. The applications of the above models to three medical image segmentation cases, including brain tumor segmentation, polyp segmentation, and heart segmentation, are presented in Sections 4, 5, and 6, respectively. Finally, the findings and future work are presented in Section 7.

2. Background

Medical imaging has been widely employed by healthcare professionals to evaluate various anatomical struc-

tures. Medical image segmentation is the process of assigning labels to individual pixels within an image, thereby converting raw images into meaningful spatial data [4]. Currently, clinicians largely perform this segmentation manually, resulting in a time-consuming process prone to both intra- and inter-observer variations [5]. The adoption of automatic segmentation methods holds significant promise, as it can enhance reproducibility and streamline clinical workflows. This is particularly relevant accounting for the demand of growing healthcare and the shortage of healthcare providers [6]. The advance of new technologies has made it possible for automatic organ segmentation [7], tumor segmentation [8], vessel segmentation [9], lesion detection and segmentation [10, 11], cardiac segmentation [12], brain segmentation [13, 14], and bone segmentation [15, 16].

Medical image segmentation is inherently influenced by the imaging modality employed. Computed tomography (CT) imaging presents challenges related to similar tissue intensities, three-dimensional data, and radiation exposure control [17]. Magnetic resonance imaging (MRI) introduces complexities in multi-contrast imaging, noises, and artifacts, as well as lengthy acquisition times [18, 19]. Ultrasound imaging, although operator-dependent and prone to speckle noises, offers real-time imaging without ionizing radiation. Understanding the distinct characteristics and challenges of each modality is crucial for selecting appropriate segmentation techniques and optimizing the accuracy of medical image analysis [20–22]. Positron emission tomography (PET) imaging, commonly used for functional studies and cancer detection, faces resolution-noise trade-offs and requires advanced algorithms for accurate segmentation when distinguishing physiological regions from pathological regions [23]. X-ray imaging faces challenges in accurate segmentation due to the inherent two-dimensional projection of three-dimensional structures [24], overlapping structures and low contrasts [25].

Historically, image segmentation can be performed by using low-level image processing methods. For examples, thresholding is a straightforward technique that involves selecting a threshold value and classifying pixels as the foreground or background based on intensity values [26]. Region-based segmentation methods focus on grouping pixels based on their spatial and intensity similarities [27]. The Watershed transform, introduced by Beucher and Serge [28], is a region-based segmentation technique that has found applications in contour detection and image segmentation.

Statistical methods have also been developed for image segmentation. K-means clustering is a widely recognized method for partitioning an image into K clusters based on pixel intensity values [29]. Active contours, introduced by Kass, Witkin, and Terzopoulos, is often referred to as “snakes” [30]. Probabilistic modelling for medical image segmentation has been presented in [31–33] where the expectation-maximisation process is adopted to model each segment as a mixture of Gaussians. The graph cut method utilises the graph theory to partition an image into distinct regions based on pixel similarities and differences. [34–37]. The level-set method, based on partial differential equations (PDE), progressively evaluates the differences among neighbouring pixels to find object boundaries, and evolves contours to delineate regions of interests [38–40, 41].

Over the past decade, the deep learning (DL) techniques stand out as the cutting-edge approach for medical image segmentation. The convolutional neural networks (CNN) are inherently suited for solving volumetric medical image segmentation tasks. The CNN can be customized by adjusting network depth and width to balance between computational efficiency and segmentation accuracy. Ensembling multiple 3D CNNs with diverse architectures has been effective in improving robustness and generalization to different medical imaging modalities [42]. Fully convolutional networks (FCN) have been successfully adapted to solve medical image segmentation tasks by fine-tuning pre-trained models or designing architectures tailored to specific challenges. In scenarios where anatomical structures exhibit varying shapes and appearances, FCN can be modified to include multiple scales and skip connections to capture both local and global information [8]. The Cov-Net has been utilised for the detection and diagnosis of COVID-19 from chest X-ray images. These models have shown promising results in accurately identifying COVID-19 cases [43]. A deep belief network-based multi-task learning approach for the diagnosis of Alzheimer's disease has gained attention due to its potential to improve accuracy and efficiency in disease diagnosis [44, 45].

The UNet [1] represents the most widely embraced variation among DL networks, featuring a U-shaped architecture with skip connections that enables the accurate delineation of objects in images [8]. The SegNet, an encoder-decoder architecture, offers adaptability to various medical imaging modalities. Its encoder can be customized to incorporate domain-specific features, such as texture and intensity variations in medical images [46]. Additionally, the decoder can be modified to handle the specific shape and structure of objects within the medical images, ensuring precise segmentation [47].

The ResUNet [2] extends the UNet architecture by introducing residual connections, which enable the network to train effectively, even with a large number of layers, thereby improving its ability to capture complex features in medical images. The integration of residual blocks in the ResUNet facilitates the training of deeper networks and enhances segmentation accuracy, making it a valuable choice for tasks demanding the precise delineation of anatomi-

cal structures in medical image analysis. Built on the ResUNet framework, the attention ResUNet [3] incorporates attention mechanisms to allow the network to selectively focus on informative regions in the input image, while suppressing noises and irrelevant features. By introducing self-attention or spatial attention modules, the attention ResUNet enhances its segmentation capabilities, particularly in scenarios in which fine details and subtle variations are critical for accurate segmentation and diagnosis.

Recently, the nnUNet automatic segmentation framework, whose self-configuration mechanism takes into consideration of both computer-hardware capabilities and dataset specific properties, has demonstrated segmentation performance that matches or closely approaches the state-of-the-art [48]. Extended models of the nnUNet have been reported in [49−51] for various medical imaging applications.

The exploration of traditional image segmentation methods has revealed both strengths and limitations in solving simpler tasks, and has exposed vulnerabilities in complex medical imaging. General segmentation techniques adapted for medical applications, such as the Watershed transform and active contours, have shown promise in specific areas with own limitations. The various domains of medical image segmentation, each with its unique challenges, highlight the complexity of this field. These challenges range from organ shape variability to vessel intricacies. In light of these challenges, the importance of the UNet and its variants becomes evident. These deep learning approaches offer the potential to overcome the limitations of traditional methods, promising more accurate and adaptable segmentation solutions to complex medical images. Exploring the UNet and its variants signifies a journey into harnessing the power of deep learning to address the intricacies of medical image segmentation. This endeavor seeks not only to understand the foundations of the UNet but also to explore its potential in overcoming the limitations of traditional methods. Ultimately, this exploration aims to advance medical image analysis, leading to improved healthcare quality and patient outcomes.

3. Methods

An overview of the deep learning models, including the UNet, Res-UNet, and attention Res-UNet, is provided in this section. The details are also given including the network architectures, filters of individual layers, connections between layers, as well as specific functional mechanisms such as attention, activation functions and normalisation.

3.1. UNet

The UNet, introduced by Ronneberger et al. in 2015 [1], is a convolutional neural network (CNN) initially designed for biomedical image segmentation but widely applied to solve various image analysis tasks. Its unique architecture includes an encoder-decoder structure with skip connections. Figure 1 a shows the general UNet architecture adopted in this paper. The UNet's architecture consists of two main components: the contracting path (encoder) and the expansive path (decoder). This design enables the UNet to capture both global and local features of the input image, making it highly effective for solving segmentation tasks.

Contracting path (encoder): The contracting path is responsible for feature extraction. The UNet model built in this paper has four encoding layers. Each encoding layer consists of 2 convolutional layers or one convolution block, each followed by batch normalisation layers for ensuring normalisation and a relu activation layer as shown in Figure 1b. The output from the convolution block is then passed through a down sampling layer with max-pooling to reduce the spatial dimensions of the feature maps. The contracting path is crucial for building a rich feature representation. After the four encoding layers, the output passes through the bottleneck layer and then the upsampling layers (decoders).

Expansive path (decoder): The expansive path aims to recover the original resolution of the image. The UNet model has four decoding layers. It comprises up-sampling and transposed convolutional layers. Importantly, skip connections connect the encoder and decoder at multiple levels. These skip connections allow the decoder to access feature maps from the contracting path, preserving spatial information and fine details.

Skip connections: Skip connections are the key innovation in UNet's architecture. They address the challenge of information loss during up-sampling. By providing shortcut connections between corresponding layers in the encoder and decoder, skip connections enable the model to combine low-level and high-level features effectively. This ensures that fine details are retained during the segmentation process.

Kernel size and number of filters: Throughout the structure, a kernel size of 3 is maintained for the convolutional layers, as this filter size is common in image segmentation. A smaller filter size captures local features, while a larger filter size captures more global features. The number of filters in the first layer is set to be 64. This is a common practice to start with a moderate number of filters and gradually increase the number of filters in deeper layers. This allows the network to learn hierarchical features.

Final fully connected convolutional layer: The output passes through a final fully connected convolutional

layer after four decoding layers. The size of kernel in the last layer depends on the number of classes (labels) present in the mask and is therefore tailored to satisfy task needs. The output from the convolutional layer passes through an activation function to produce the final output. The final activation function also depends on the number of labels in the output. The final Kernel size and the activation layer are mentioned for each task in the following sections.

The design of the UNet makes it particularly effective for solving tasks where precise localization and detailed segmentation are required, such as medical image segmentation.

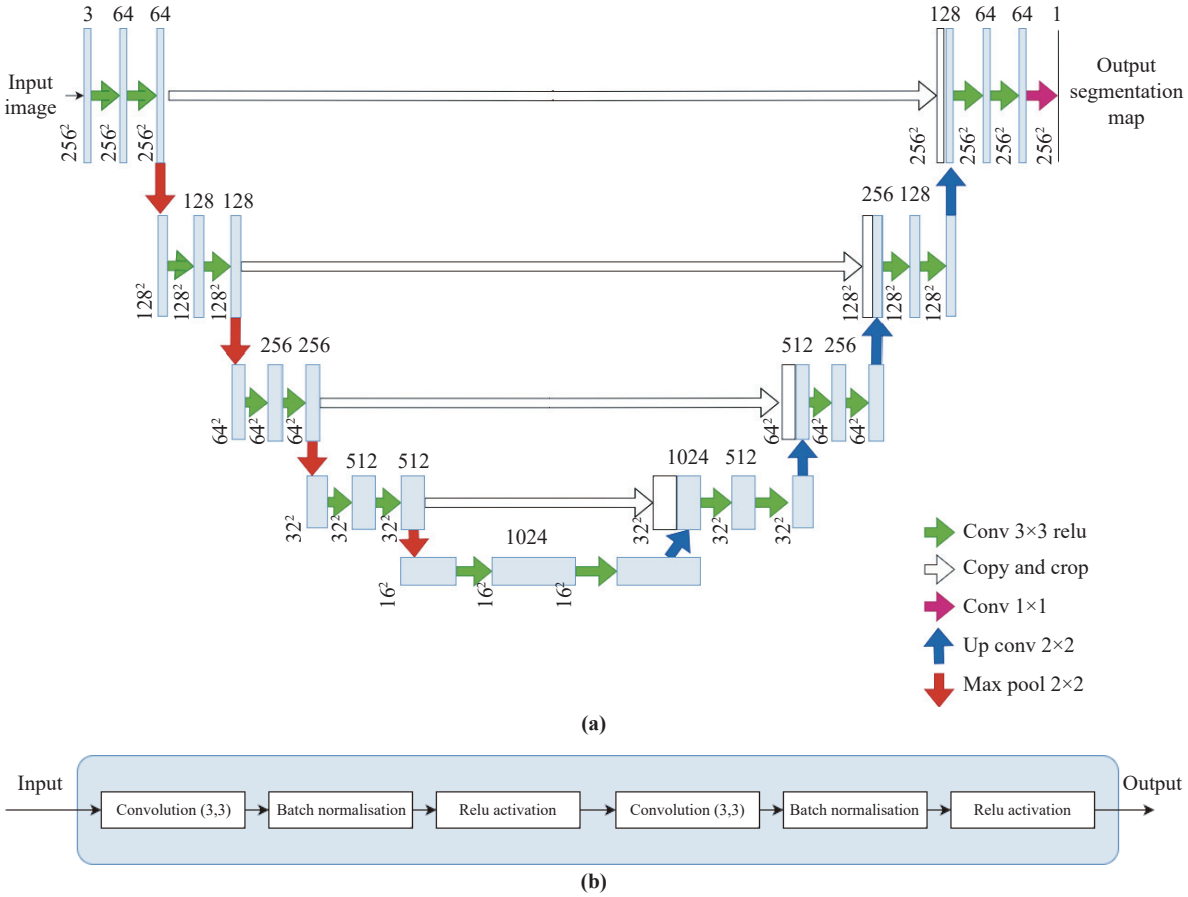

Figure 1. UNet. **(a)** Network architecture. **(b)** Details of the convolution block.

3.2. Res-UNet

The Res-UNet is an extension of the UNet that incorporates residual connections. Residual connections have been introduced in the context of residual networks (ResNets) [2] to address the vanishing gradient problem in deep networks. The Res-UNet combines the strengths of the UNet with the benefits of residual connections. The convolution block in the UNet is replaced here with residual blocks, and this introduces an addition layer between the input at each block and the output from the last 3x3 convolutional block.

Residual connections: The Res-UNet incorporates residual connections between layers. These connections allow gradients to flow more easily during training, enabling the training of deeper networks without suffering from vanishing gradients.

Enhanced information flow: The use of residual connections enhances the flow of information through the network, enabling it to capture long-range dependencies and complex structures in medical images.

The Res-UNet model adopted in this paper has four encoding and four decoding layers. The overall architectures of the Res-UNet model and the residual convolutional block are provided in Figure 2a and 2b, respectively. The Res-UNet is known for its ability to handle deeper networks, which can be advantageous for capturing intricate details in medical images.

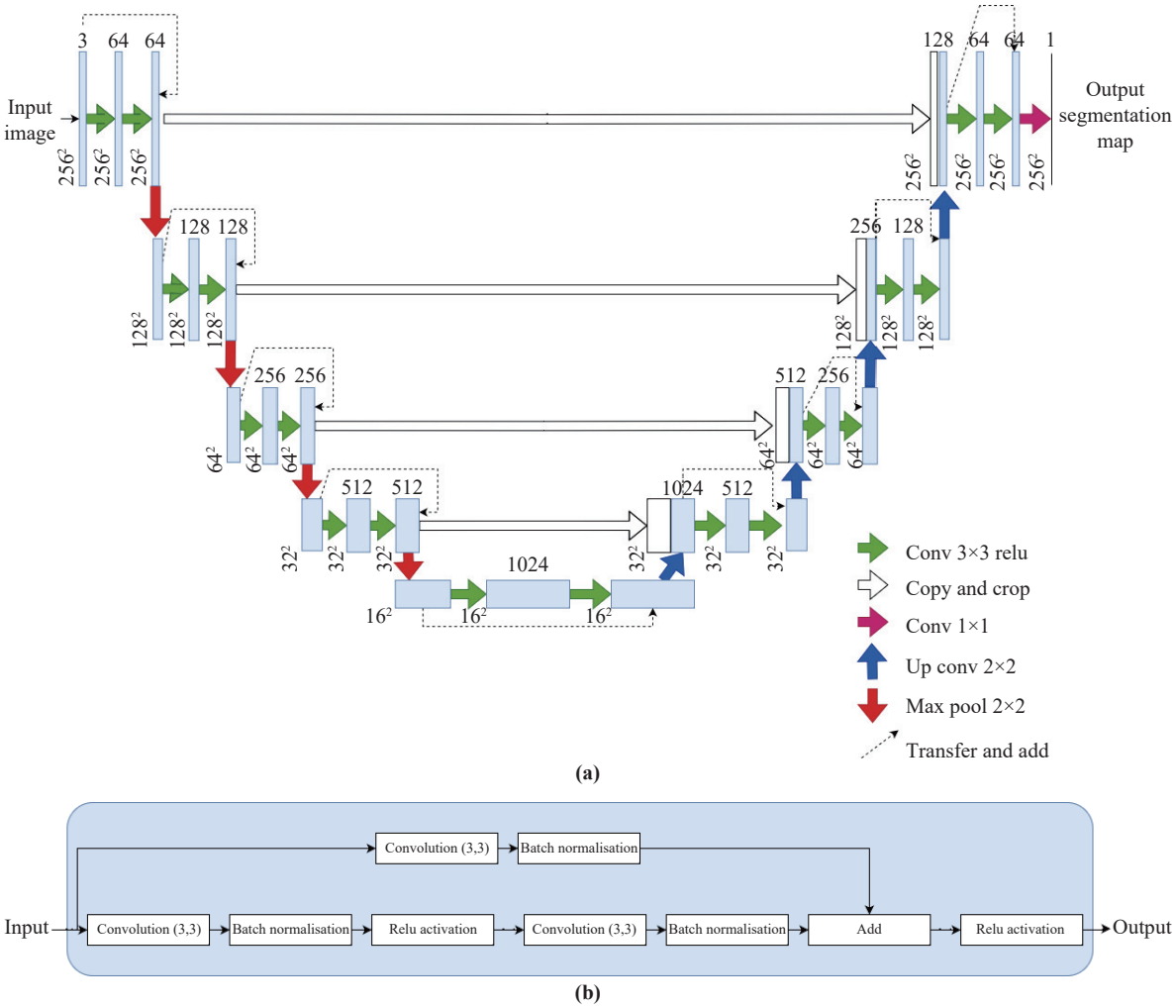

Figure 2. Res-UNet. **(a)** Network architecture; **(b)** Details of the residual convolution block.

3.3. Attention Res-UNet

As illustrated in Figure 3, the attention Res-UNet model [3] is built on the Res-UNet architecture, which introduces spatial attention mechanisms. This is achieved through the gating signal which brings output from the lower layer to match the same dimension as the current layer, and an attention block which combines information from two sources: the input feature map (x) and the gating signal (gating) to compute attention weights. The specific attention mechanism used here is the spatial attention, where the attention weights are computed based on relationships between different spatial regions in the input feature map. This allows the model to focus on salient parts of the image while suppressing irrelevant regions. The attention block consists of the following components. The query convolution applies convolutions on the input feature map which is later transformed into a query that captures spatial interdependencies. The key convolution transforms the gating signal into keys to highlight relevant spatial regions. The matrix multiplication combine the query and keys to determine spatial attention weights. The softmax normalises the attention weights to values between 0 and 1; The value convolution transforms the input feature map into values representing the original features. The weighted sum multiplies the attention weights with the values and sums all resultant values up.

The attention mechanism enables the network to focus on salient regions of the input, improving its ability to differentiate between important and less important features. The key steps to implement the two blocks are explained below.

Gating Signal:

The gating signal is a subnetwork or a set of operations employed to modulate the flow of information in an attention mechanism. In this specific implementation, the gating signal is generated as follows:

• **Convolutional layer:** A convolutional layer is used to transform the input feature into a format compatible with the requirement of the attention mechanism. It adjusts the feature dimensionality if necessary.

• **Batch normalization (optional):** An optional batch normalization layer is applied to ensure that the output of the convolutional layer is well-scaled and centered, thereby aiding in stabilizing the training process.

• **ReLU activation:** The ReLU activation function introduces non-linearity to the gating signal, helping capture complex patterns and relationships in the data.

attention block: The attention block is a critical part of attention mechanisms employed in neural networks. Its primary purpose is to combine information from two sources—the input feature map (x) and the gating signal (*gating*). The breakdown of the functionality of the attention block is given below.

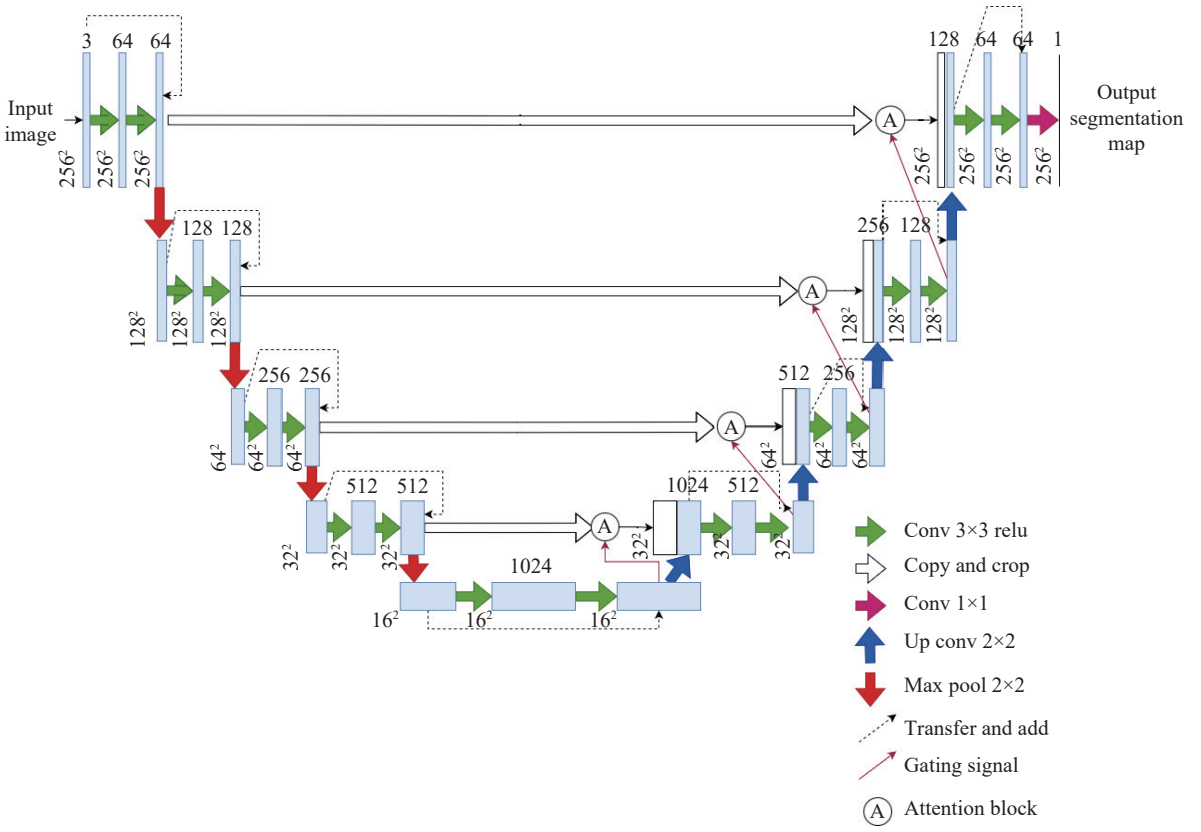

Figure 3. Proposed attention Res-UNet Architecture

• **Spatial transformation (Theta_x):** The input feature map (x) undergoes spatial transformation using convolutional operations. This transformation ensures that the feature map aligns with the dimension of the gating signal.

• **Gating signal transformation (Phi_g):** Similarly, the gating signal is subjected to transformation via convolutional operations to ensure appropriate spatial dimensions.

• **Combining information:** The transformed gating signal (*Phi_g*) and the spatially transformed input feature map (*Theta_x*) are combined to capture relationships between different parts of the input.

• **Activation (ReLU):** The ReLU activation function is applied to the combined information, introducing non-linearity and enabling the capture of complex relationships.

• **Psi and sigmoid activation:** The combined information is further processed to produce attention weights (*Psi*) using convolutional layers and a sigmoid activation function. The sigmoid activation ensures that the attention weights are within the range of 0 to 1, indicating the degree of attention assigned to each spatial location.

• **Upsampling Psi:** The attention weights are upsampled to match the spatial dimensions of the original input feature map, ensuring alignment with the input.

• **Multiplication (attention operation):** The attention weights are multiplied element-wisely with the original input feature map (x). This operation effectively directs attention to specific spatial locations in the feature map based on the computed attention weights.

• **Result and batch normalization:** The final result is obtained by applying additional convolutional layers and optional batch normalization, ensuring that the output is appropriately processed.

The gating signal prepares a modulating signal that influences the attention mechanism in the attention block. The attention block computes attention weights with the focus on relevant spatial regions of the input feature map, which is particularly useful in addressing tasks requiring the capture of fine-grained details such as image segmenta-

tion or object detection. The attention mechanism aids the network in prioritizing and weighting different spatial locations in the feature map, ultimately enhancing the performance.

3.4. Evaluation Methods

The following metrics are adopted to evaluate the performance of the models.

Execution time: Execution time is recorded for the training of each model to understand how long a model takes to converge. This is implemented using the datetime library in Python.

Validation loss over epochs: The change in the validation loss over the training period gives a glance on model convergence. Model convergence graphs show that how well the model is trained and how efficient a model converges. These graphs show the lowest loss required for the validation data, and the fluctuation in the loss which evaluates model stability. The graphs provide an initial basis of comparisons between different models.

The dice similarity coefficient: Also known as the sørensen-dice coefficient, which is a metric used to quantify the similarity or overlap between two sets or groups. In the context of image segmentation and binary classification, the dice coefficient is commonly employed to evaluate the similarity between two binary masks or regions of interest (ROIs).

Formally, the dice similarity coefficient (DSC) is defined as

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

where:

A is the first set or binary mask (e.g., the predicted segmentation mask);
B is the second set or binary mask (e.g., the ground truth or reference mask);
$|\cdot|$ denotes the cardinality of a set, i.e., the number of elements in the set;
$\cap$ denotes the intersection operation, which yields the common elements between sets A and B.

The dice coefficient produces a value between 0 and 1, where

• $DSC = 0$ indicates no overlap or dissimilarity between the two sets. It means that there is no commonality between the predicted and reference masks.

• $DSC = 1$ indicates perfect overlap or similarity between the two sets. It means that the predicted mask perfectly matches the reference mask.

In the context of image segmentation, the dice coefficient is a valuable metric because it measures the agreement between the segmented region and the ground truth. It quantifies how well the segmentation result matches the true region of interest. Higher DSC values indicate better segmentation performances.

Intersection over union (IoU) or Jaccard index: The IoU measures the overlap between the predicted segmentation mask (A) and the ground truth mask (B). It is calculated as the intersection of the two masks divided by their union. The higher the IoU, the better the segmentation accuracy.

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

where:

A is the predicted mask;
B is the ground truth mask.

In this formula, $|A \cap B|$ denotes the cardinality of the intersection of sets A and B, and $|A \cup B|$ represents the cardinality of their union. The IoU quantifies the extent to which the predicted mask and the ground truth mask overlap with each other, providing a valuable measure of the segmentation accuracy. The implementation of the Jaccard index using Python is given below. **Confusion matrix:** A confusion matrix provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions. It is useful for understanding the model's performance on different classes or categories within the segmentation. This is implemented using the confusion_matrix function in Python's sklearn.

Precision: Precision assesses the accuracy of an algorithm in correctly identifying relevant pixels or regions. It's the ratio of true positive pixels (correctly segmented) to all pixels identified as positive by the algorithm. High precision indicates that it's usually correct when the algorithm marks a pixel or region as a part of the target object. In medical image segmentation, high precision means that it's likely to be accurate when the algorithm identifies an area as a specific organ or structure, thereby reducing the false positive.

Recall: Recall, also called the sensitivity or true positive rate, gauges an algorithm's capacity to accurately iden-

tify all relevant pixels or regions in an image. It's the ratio of true positive pixels to the total pixels constituting the actual target object or region in the ground truth. A high recall value signifies that the algorithm excels at locating and encompassing the genuine target object or region. In medical image segmentation, high recall means that the algorithm effectively identifies and includes most relevant anatomical structures, reducing the likelihood of the false negatives

4. Brain Tumor Segmentation

The task of Brain tumor segmentation involves the process of identifying and delineating the boundaries of brain tumors in medical images, specifically in brain MRI scans. The goal of this segmentation task is to automatically outline the shape and extent of lower-grade gliomas (LGG) within the brain images.

4.1. Pre-processing

The dataset used in this study was obtained from Kaggle [52] and was originally sourced from The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG) [53]. The dataset includes brain MR images that are accompanied by manually created FLAIR abnormality segmentation masks. The dataset contains MRI FLAIR image data of 110 patients. Each MRI image is a RGB image with three channels, and each mask is a 2D black and white image.

The dataset originally contains 1200 patient images and masks, with 420 masks indicating the presence of tumors. To focus the model on tumor segmentation, images without tumor annotations are removed. The dataset is then split into training, testing, and validation sets using an 8:1:1 ratio. To handle this data efficiently, a data pre-processing step is employed — a crucial tool in deep learning, particularly for large datasets that don't fit in memory. In this step, the data is processed in smaller batches during training, effectively managing computational resources and ensuring real-time preprocessing during model training. Details of the pre-processing steps are listed below.

1) **Image resizing**: Images in the dataset are resized to a standard 256 by 256 pixel dimension to ensure compatibility with neural network architectures. This choice balances between preserving important details (i.e. which smaller size might lose), and avoiding unnecessary noise (i.e. which larger size could be introduced).

2) **Standardization**: Both image and mask data are standardized by adjusting their pixel values to have a mean of 0 and a standard deviation of 1. This uniform scaling simplifies data for deep learning models, promoting convergence and training stability.

3) **Normalisation of mask images**: The mask images, initially with binary values (0 for background, 1 for the mask), have their values become floating-point during resizing. To prepare images for model training, their dimensions are expanded by one to (256x256x1), followed by a thresholding operation. Pixel values greater than 0 are set to be 1 (indicating a tumor), while values equal to or less than 0 are set to be 0 (representing the background), thereby maintaining binary values suitability for training.

4.2. Model Training

4.2.1. Loss Function

The **Binary focal loss** (BFL) is a specialized loss function used in binary classification, particularly when dealing with imbalanced datasets or cases where certain classes are of greater interest than others. The BFL is designed to address the problem of class imbalance with the focus on improving the learning of the minority class. Formally, the BFL is defined as follows:

$$BFL = -(1 - p_t)^{\gamma} \cdot \log(p_t) \tag{2}$$

where

p_t represents the predicted probability of the true class label;
γ is a tunable hyperparameter known as the focusing parameter;
$\log(\cdot)$ is the natural logarithm.

The BFL has the following key characteristics.

• It introduces the focusing parameter γ to control the degree of importance assigned to different examples. A higher γ emphasizes the training on hard and misclassified examples, while a lower γ makes the loss less sensitive to those examples.

• When $\gamma = 0$, the BFL reduces to the standard binary cross-entropy loss.

• The term $(1 - p_t)^{\gamma}$ is a modulating factor that reduces the loss for well-classified examples (p_t close to 1) and

increases the loss for misclassified examples (p_t close to 0).

• The BFL helps the model focus more on the minority class, which is especially useful for imbalanced datasets where the majority class dominates.

• The BFL encourages the model to learn better representations for challenging examples, potentially improving overall classification performances.

• The loss is applied independently to each example in a batch of data during training.

The BFL is a valuable tool for addressing class imbalance and improving the training of models for imbalanced binary classification. By introducing the focusing parameter, practitioners are allowed to fine-tune the loss function according to the specific characteristics of the dataset and the importance of different classes.

4.2.2. Model Design Choices

UNet: The UNet model has input shape of (256,256,3) for the RGB images and an output layer of shape (256,256, 1) for the mask output. The final output layer consists of 1x1 convolutional layers followed by batch normalization and sigmoid activation. These layers produce the segmentation mask, where each pixel is classified as either a part of the object or background. Sigmoid activation is used for binary segmentation. The model has a total of 31,402,501 parameters with 31,390,723 of them trainable.

Res-UNet: The Res-UNet model takes RGB images with an input shape of (256,256, 3) and produces a mask output with an output layer of shape (256,256, 1). The last layer of the model comprises 1x1 convolutional layers, followed by batch normalization and sigmoid activation.

attention Res-UNet: The attention Res-UNet follows the same input and output configuration as the previous models due to the same input and output image and mask specifications.

4.2.3. Callbacks

Three callbacks are assigned to the models.

1) Early stopping callback (early stopping): The early stopping callback monitors validation loss during training. If there's no improvement (decrease) for 20 epochs, training stops early to prevent overfitting and save time, thereby ensuring that the model does not learn noises or deviate from the optimal solution.

2) Reduce LROn Plateau callback: The reduce LROn plateau callback is used to optimize the model's training by lowering the learning rate when the validation loss reaches a plateau or stops improving, aiding the model in fine-tuning and avoiding the local minima. The callback monitors "val_loss" with a "min" mode, providing informative updates (verbose = 1), and adjusts the learning rate if there's no improvement in validation loss for 10 consecutive epochs (patience = 10) by a factor of 0.2 (reduced to 20% of its previous value), thereby ensuring meaningful improvement with a "min_delta" parameter set at 0.0001.

3) Check pointer: The check pointer is specified to save weights of the trained model only when the validation loss improves.

4.2.4. Model Compilation and Fitting

The models are compiled using the the Adam optimizer with an initial learning rate of "1e-5". Multiple initial learning rates are tested, where higher rates cause divergence and lower rates slow down training. Two compilations are done for each model, where the one uses the dice-coefficient as the loss function and the other uses the BFL. Training is performed on both the training and validation data for 100 epochs initially.

4.3. Results

The model training times and epochs are listed in Table 1. The model differences in execution times indicate varying computational resource requirements during training. Notably, the attention Res-UNet emerges as the model with the longest training duration. This extended duration could be attributed to the model's complexity which necessitates additional time for convergence.

Table 1 Execution time and epochs of trained models

Models	Execution Time	Epoch
UNet	34 min 20 sec	69
Res-UNet	42 min 1 sec	89
attention Res-UNet	1 hr 1 min	100

Regarding the raining behavior, the number of epochs completed by each model offers insights into their

respective convergence behaviors. The UNet model exhibits a comparatively lower number of epochs, implying relatively swift convergence. This is indicative of a particularly efficient training process. Conversely, the Res-UNet and attention Res-UNet need more extensive training, implying that potentially more intricate model architectures or more requirements are needed to achieve convergence after extended training periods.

Furthermore, it is worth noting that some models end training prematurely due to a lack of improvement in validation loss, as evidenced by lower epoch counts. This highlights the consideration of the early stopping strategy, which is a common technique used to curtail training and prevent overfitting. This observation raises the need for discussions on optimizing model performances and making thoughtful decisions about the resource allocation during training. The sub-figures in Figure 4 depict the evolution of the BFL over epochs of training and validation data for three different models. These results offer insights into how these models perform during the training process.

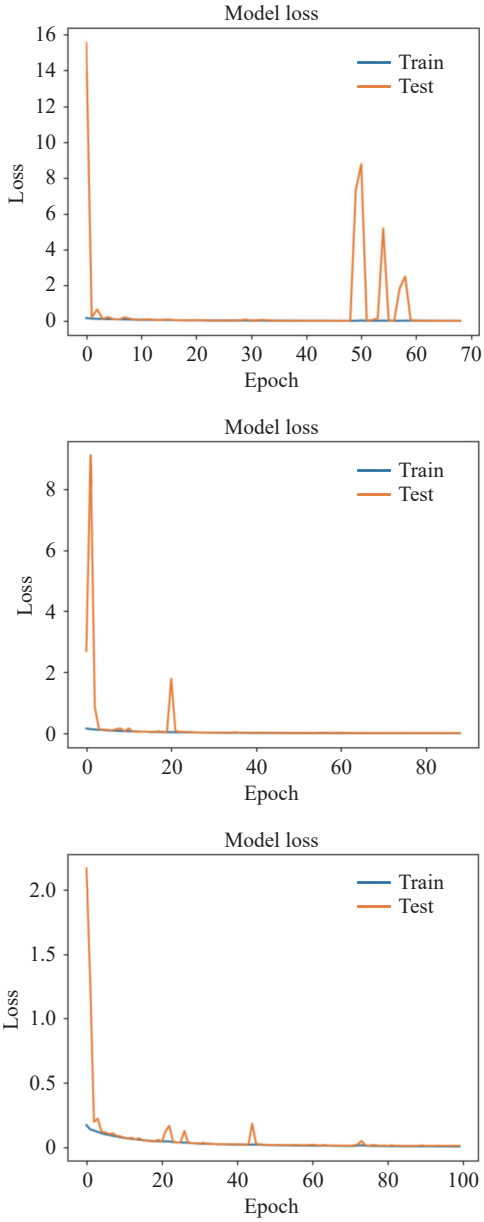

Figure 4. Changes in binary focal loss for each models, from top to bottom, UNet, Res-UNet and Attendion Res-UNet.

1) **Initial validation loss:** Initial validation losses vary among the models. The UNet starts with a high initial loss (around 15), indicating initial difficulties in accurate prediction. In contrast, the Res-UNet begins with a lower loss (around 8), while the attention Res-UNet starts with an even lower loss (approximately 2), suggesting that the two models make relatively better prediction from the start.

2) **Early epoch performance:** All three models exhibit a rapid decrease in the validation loss within the first ten epochs. This implies that models quickly learn to capture relevant patterns in the data and improve prediction performances during this early training phase.

3) **Stability in training:** During training, all models maintain generally low validation losses, with some fluctuations. UNet exhibits significant fluctuations towards training's end, suggesting sensitivity to data variations. In contrast, Res-UNet shows minor early fluctuations but stabilizes. attention Res-UNet also experiences initial fluctuations, but they are much smaller than in the other models.

4) **Comparison of model performance:** The UNet quickly reduces the validation loss at the start, but has higher fluctuations later. The Res-UNet starts with a moderate loss, experiences some early fluctuations, and finally stabilizes. The attention Res-UNet consistently performs well from the beginning with the minimal fluctuation.

Overall, these results highlight trade-offs between the rapid initial learning and the model stability. The UNet learns quickly but exhibits greater instability, while the Res-UNet and attention Res-UNet provide more consistent and reliable prediction performances. Table 2 provides performance metrics for the UNet, Res-UNet, and attention Res-UNet, when applied to test data.

Table 2 Performance Metrics for UNet, Res-UNet, and attention Res-UNet on test data

Model	Focal Loss	Accuracy	Precision	Recall	Dice	IoU
UNet	0.0169	0.987	0.852	0.623	0.72	0.563
Res-UNet	0.0062	**0.996**	**0.923**	0.939	**0.931**	**0.870**
attention Res-UNet	**0.0055**	**0.996**	0.902	**0.946**	0.923	0.858

1) **Focal loss:** All the models achieve low focal losses with the Res-UNet and attention Res-UNet outperforming the UNet. The attention Res-UNet achieves the lowest focal loss, highlighting its proficiency in addressing the problem of class imbalance. This means that the variants perform better at focusing on hard-to-classify pixels, which is the tumor class.

2) **Accuracy:** The Res-UNet and attention Res-UNet exhibit impressive accuracies of approximately 99.6%, surpassing the 98.7% accuracy of the UNet. Both Res-UNet and attention Res-UNet excel in pixel-level classification.

3) **Precision and recall:** The Res-UNet demonstrates superior precision, indicating accurate positive pixel classification with the minimal false positive. The UNet and attention Res-UNet exhibit slightly lower precision values. Conversely, the attention Res-UNet achieves the highest recall, suggesting its effectiveness in capturing a larger proportion of true positives.

4) **Dice coefficient:** The Res-UNet achieves the highest dice coefficient at approximately 0.931, signifying accurate spatial prediction performances. The UNet and attention Res-UNet yield slightly lower dice coefficients, but maintain strong performances.

5) **Intersection over union (IoU):** The Res-UNet achieves the highest IoU of approximately 0.870, indicating the superior spatial overlap. The UNet and attention Res-UNet record slightly lower IoU values, though they continue to deliver commendable results in this aspect.

In summary, the Res-UNet and attention Res-UNet consistently outperform the UNet across multiple performance metrics, underscoring their superior performances in image segmentation of the test data. The Res-UNet excels in the precision, dice coefficient, and IoU, while the attention Res-UNet achieves the highest recall.

4.4. Discussions

Figure 5 shows four examples with given images and ground truth masks followed by the prediction from the three models. The four examples are chosen as they represent different types of results observed in the whole test prediction.

The UNet exhibits sensitivity to tumor features and shows promising performances in identifying likely tumor locations, but tends to misclassify tumor pixels as the background, leading to false negatives. The UNet also mistakenly classifies the background pixels as tumors, causing false positives and impacting precision. The Res-UNet and attention Res-UNet, on the other hand, deliver highly accurate predictions (thereby capturing fine details and maintaining a balance between sensitivity and specificity), but occasionally overestimate tumor presence.

The UNet and its variants perform adequately in most cases, but struggle when tumors are very small or have complex boundaries. The UNet also faces challenges of class imbalance, resulting in misclassification and poor recall. The Res-UNet and attention Res-UNet mitigate these limitations, successfully locating tumors in challenging conditions and reducing misclassifications significantly. The attention Res-UNet excels in handling the problem of class imbalance.

Despite variations in performances, all models achieve high accuracy scores. When tumors are misclassified due to the relatively small tumor size compared to the background, accuracy may not be a reliable metric as it can

remain high, making it unreliable for assessing model performances.

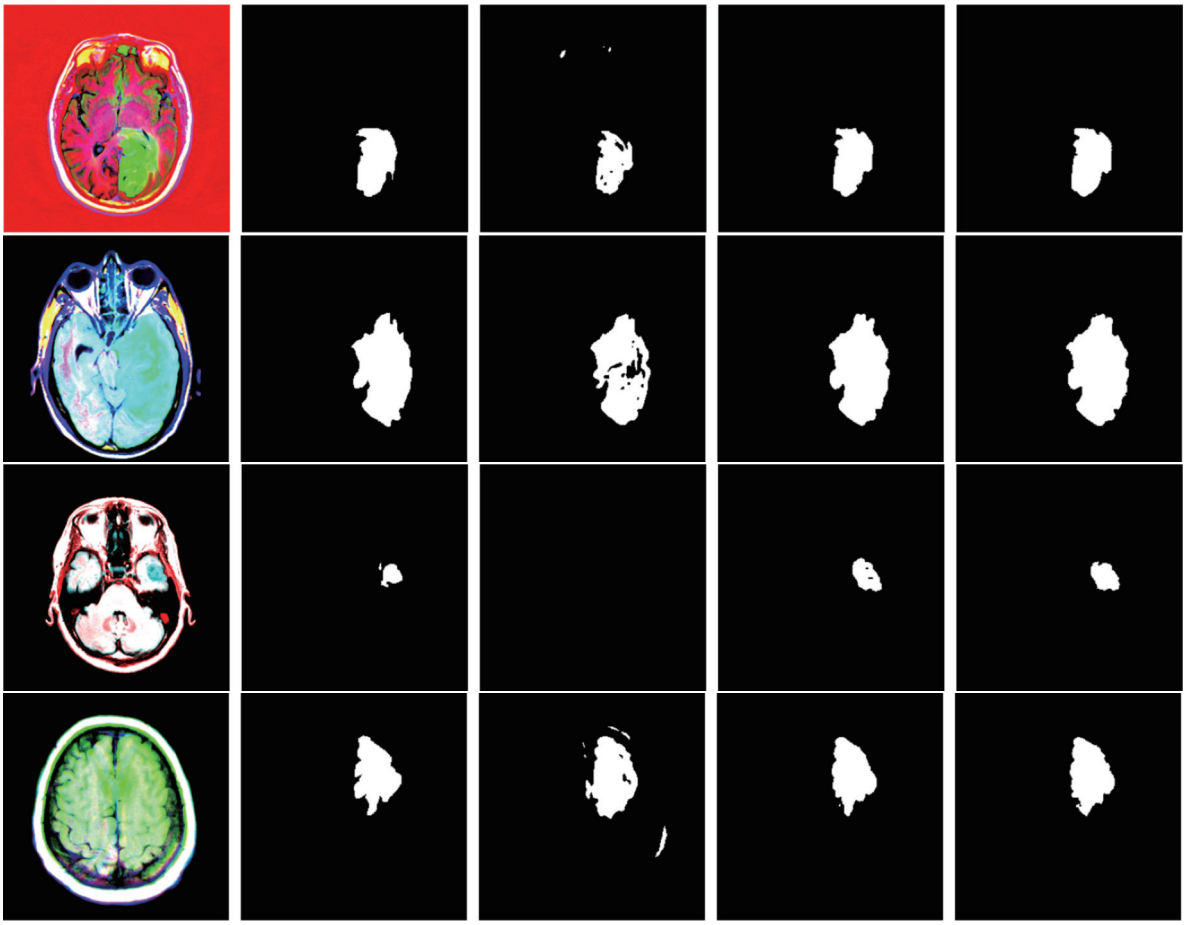

Figure 5. Segmentation results by the three models for four different examples, from left to right are the input images, ground-truth, segmentation results by UNet, Res-UNet and attention Res-UNet.

5. Polyp Segmentation

Polyp segmentation refers to the process of identifying and delineating the boundaries of polyps in medical images, particularly in the context of medical imaging, endoscopy, and colonoscopy. The goal of this segmentation task is to automatically outline the shape and extent of polyps from colonoscopy images.

5.1. Pre-processing

The CVC-ClinicDB dataset [54], whcih is established in the Hospital Clinic of Barcelona, Spain, is utilized in this segmentation task, where featuring frames are extracted from colonoscopy videos showcasing polyps. The dataset is generated from 23 different video studies of white light standard colonoscopy interventions. The CVC-ClinicDB database contains 612 polyp images with a size of 576 × 768. It includes corresponding ground truth masks outlining polyp regions. The dataset consists of two main types of images: the original colonoscopy frames at 'original/frame number.tiff' and the corresponding polyp masks at 'ground truth/frame number.tiff'.

A pandas dataframe is employed to manage images and mask paths. The dataframe is used to split the data into training, testing, and validation sets in an 8:1:1 ratio. A dataset generator processes images and masks in the training and validation data one by one, using a 'tf_parse()' function to read, resize, and preprocess for compatibility with the program's requirements. The pre-processing steps are listed below.

1) **Reading the image:** The function first reads the image from the file path x using OpenCV (cv2.imread). This reads the image as it is in its original form.

2) **Resizing the image:** After reading, the image is resized to a fixed size of 256 × 256 pixels using OpenCV's cv2.resize function. This resizing ensures that all images have the same dimensions, which is typically necessary for training deep learning models.

3) **Normalizing the image:** The pixel values of the resized image are scaled to a range between 0 and 1. This is

done by dividing all pixel values by 255.0. Normalizing the pixel values helps the deep learning model learn more effectively.

5.2. Model Training

5.2.1. Loss Function

The BFL is used as the loss function for the models. The masks in the problem are binary (label and background), followimg the similar design as the brain tumor problem.

5.2.2. Model Design Choices

The model design choices for the **UNet**, **Res-UNet** and **Attention Res-UNet** for Polyp segmentation are similar to the the models used for brain tumor segmentation. The two problems, even though crucial in their own ways to the medical community, share the same configuration by the fact that they involve creating binary segmentation masks from RGB images. Hence, the input shape for the images in both problems is (256,256, 3), while the output shape is (256,256, 1). This does not require a change in the model architectures.

5.2.3. Callbacks

The callback used for this problem is the early stopping, The reducing learning rate and checkpointers are the same as the ones mentioned in the previous problem.

5.2.4. Model Compiling and Fitting

Models are compiled and fitted for 100 epochs using the Adam optimizer with an initial learning rate 1e-5.

5.3. Results

Model training times and epochs are listed in Table 3.

Table 3 Execution time and epochs of trained models for Polyp Segmentation

Models	Execution Time	Epoch
UNet	51 min 14 sec	73
Res-UNet	45 min 12 sec	63
Attention Res-UNet	56 min 45 sec	62

The attention Res-UNet has the longest training duration of 56 minutes and 45 seconds, likely due to its complex architecture. The UNet shows efficient training, completing in 73 epochs. The Res-UNet and attention Res-UNet require more extensive training durations of 45 minutes, 12 seconds and 62 epochs. Some models end training early due to no improvement in the validation loss, highlighting the importance of early stopping strategies in preventing overfitting. This emphasizes the need for the optimal model performance as well as resource allocation decisions during training.

Figure 6 depict the evolution of the BFL over epochs for validation data of three different models. These results offer insights into how these models perform during the training process.

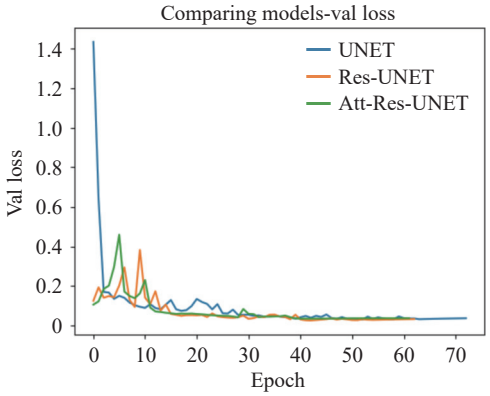

Figure 6. Convergence for trained models on Polyp Segmentation.

UNet model:

The UNet model initiates training with a high validation loss, approximately 1.4, primarily because its initial

weights are far from the optimal ones. Note that within the initial ten epochs, the UNet model experiences a rapid decrease in the validation loss, which is a common case during the early training stages of many neural networks. This reduction reflects the model's improvement in fitting the training data as it adjusts weights through techniques like backpropagation and stochastic gradient descent (SGD). After this initial phase, the UNet model maintains a relatively low and stable loss for the remaining epochs, albeit with some minor fluctuations. These fluctuations are likely attributable to inherent data noises and the stochastic nature of the optimization process.

Res-UNet and attention Res-UNet models:

Both Res-UNet and attention Res-UNet models are trained with a low initial validation loss, roughly 0.2, suggesting potential pretraining or initialization and reaching to a reasonable starting point. In the initial 15 epochs, both models experience fluctuations in the loss. This is common during initial training phases as both models adapt to the data and fine-tuned weights, possibly indicating sensitivity to the initial configuration or data noises. As training progresses, both models achieve stable loss values, signifying that models have reached a consistent and relatively optimal solution compared to the UNet within this timeframe. Eventually, all models reach the minimum loss of approximately 10%, demonstrating similar performance levels in minimizing the loss on the validation data, despite differences exist in the convergence speed and early fluctuations.

In summary, these results indicate that the UNet initiates training with a higher loss but converges swiftly. In contrast, the Res-UNet and attention Res-UNet begin with lower losses but may show more early training fluctuations. Nevertheless, all models ultimately achieve a similar minimum loss, showcasing their ability to capture crucial data features and make accurate predictions. Table 4 provides performance metrics for the UNet, Res-UNet, and attention Res-UNet when applied to test data.

Table 4 Performance Metrics for the UNet, Res-UNet, and attention Res-UNet on test data

Model	Focal Loss	Accuracy	Precision	Recall	Dice	IoU
UNet	0.0387	0.968	0.913	0.733	0.813	0.686
Res-UNet	**0.0369**	**0.971**	**0.925**	0.766	**0.838**	**0.721**
Attention Res-UNet	0.0394	0.969	0.881	**0.788**	0.832	0.712

1) **Focal loss:** All the models achieve low focal loss values, with the Res-UNet and attention Res-UNet outperforming the UNet. The Res-UNet achieves the lowest focal loss, highlighting its proficiency in addressing the problem of class imbalance. This means that the variants perform better at focusing on hard-to-classify pixels, which is the tumor class.

2) **Accuracy:** The Res-UNet and attention Res-UNet exhibit impressive accuracies, approximately 99.6%, surpassing the 98.7% accuracy of the UNet. Both Res-UNet and attention Res-UNet excel in pixel-level classification.

3) **Precision and recall:** The Res-UNet demonstrates superior precision, indicating accurate positive pixel classification with the minimal false positives. The UNet and attention Res-UNet exhibit slightly lower precision values. Conversely, the attention Res-UNet achieves the highest recall, suggesting its effectiveness in capturing a larger proportion of true positives.

4) **Dice coefficient:** The Res-UNet achieves the highest dice coefficient at approximately 0.931, signifying accurate spatial predictions. The UNet and attention Res-UNet yield slightly lower dice coefficients, but maintain strong performances.

5) **Intersection over union (IoU):** The Res-UNet achieves the highest IoU of approximately 0.870, indicating superior spatial overlap. The UNet and attention Res-UNet record slightly lower IoU values, though they continue to deliver commendable results in this aspect.

In summary, the Res-UNet and attention Res-UNet consistently outperform the UNet across multiple performance metrics, underscoring their superior performances in image segmentation on the test data. The Res-UNet excels in precision, dice coefficient, and IoU, while the attention Res-UNet achieves the highest recall. Figure 7 shows four examples with the given image and ground truth mask followed by predictions from the three models.

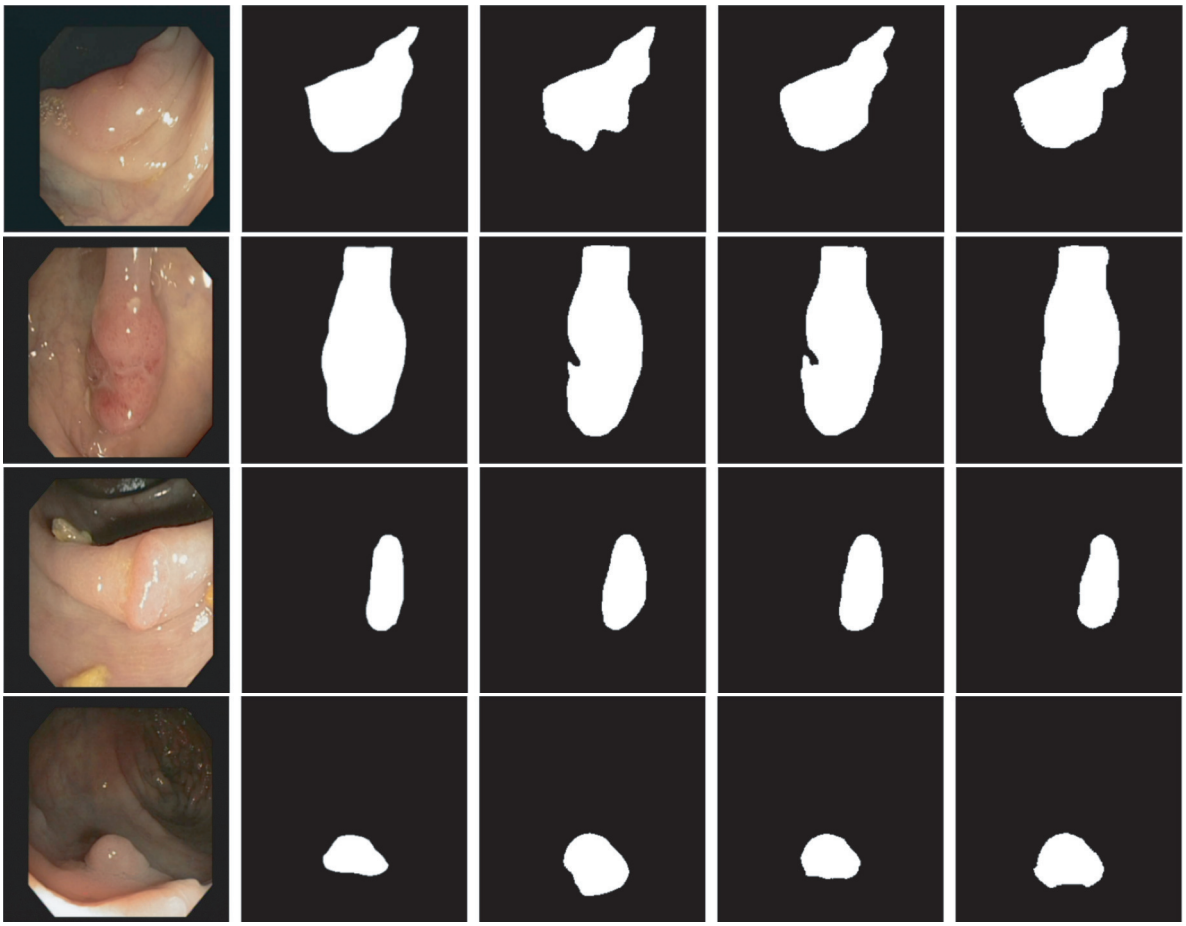

Figure 7. Segmentation results by the three models for four different examples, from left to right are the input images, ground-truth, segmentation results by UNet, Res-UNet and Attention Res-UNet.

5.3.1. Discussion

Polyp segmentation presents challenges due to the irregular and random sizing of polyps as well as limiting generalization, which can be exacerbated by data limitations. All trained models show average segmentation results. The UNet, as the base model, trains and converges quickly, especially benefiting from the less imbalanced nature of polyp scans compared to brain MRI masks.

The UNet exhibits lower performances in predicting the target class (reflecting its sensitivity to polyp features), and struggles with class imbalance. The UNet occasionally misclassifies some polyp pixels as the background (false negatives) and the background pixels as polyps (false positives), impacting both sensitivity and precision. The low true positive score in the confusion matrix underscores these challenges in accurate polyp detection.

In contrast, the Res-UNet and attention Res-UNet perform consistently (reflecting their performances in brain tumor segmentation), and excel in capturing intricate edge boundaries and maintaining the accuracy with small ground truth masks. There are rare instances of slight overestimation of polyp presence. Misclassifications are minor and have the minimal impact. Accounting for its low recall score, we know that the attention Res-UNet is better at predicting true positives than other models.

6. Heart Segmentation

The third task involves the multi-label segmentation of cardiac structures in medical images, and specifically, the target of the left ventricle (LV), right ventricle (RV), and myocardium. Accurate segmentation of the LV is essential for assessing its size and function, while RV segmentation aids in diagnosing cardiac conditions. Furthermore, precise myocardium segmentation provides insights into its thickness and function, offering indicators of heart health and potential issues.

6.1. Data Pre-processing

The "Automatic Cardiac Diagnosis Challenge" (ACDC) [55] dataset is used in this segmentation task. This is

the largest publicly available and fully annotated dataset for cardiac MRI (CMR) evaluation purposes. The dataset encompasses data from 150 CMRI recordings which are stored in a 4D "nifti" format, preserving the original image resolution and primarily containing the whole short-axis slices of the heart. This specifies the diastolic and systolic phases of the cardiac cycle. The MRI images are in grayscale, while the mask images employ a 0 to 3 scale, with 0 representing the background, 1 corresponding to the RV cavity, 2 representing the myocardium, and 3 corresponding to the LV cavity.

The preprocessing steps involve creating a dataframe to record image and mask volumes, reading them using the "nibabel" library, and iterating through slices in the third dimension of both the image and mask volumes. Each slice is cropped using a custom "crop" function with most images having a minimum dimension less than 150, and this leads to a final size of (128,128) to avoid introducing noises or unreliable information.

Mask images, with pixel values ranging from 0 to 3 (representing the labels and background), are converted to **one-hot encoding** by increasing the dimensionality to 4. This is a crucial step for generating multi-label loss functions and accurate predictions. For instance, a pixel value of 0 becomes (0, 0, 0, 0), while 3 becomes (0, 0, 0, 1).

MRI pixel values, with the maximum of 3049, are **normalized** to a range of 0 to 1, making them compatible with neural networks. These preprocessing steps are essential for preparing the data for model training.

6.2. Model Training

6.2.1. Loss Function

The categorical focal loss is used as the loss function in the multi label segmentation task.

The **categorical focal cross-entropy** combines the concepts of the categorical cross-entropy and focal loss to create a loss function suitable for addressing multi-class segmentation tasks with class imbalance. It introduces the focal loss component into the standard categorical cross-entropy. This helps the model focus on harder-to-classify pixels when handling imbalanced datasets.

$$CFC(y,p) = -\sum_{i=1}^{N} \alpha_i \cdot (1-p_i)^{\gamma} \cdot y_i \cdot \log(p_i) \tag{3}$$

In summary, the categorical focal cross-entropy is a loss function that blends the properties of the categorical cross-entropy and focal loss to improve the training of models in imbalanced multi-class segmentation tasks. The categorical focal cross-entropy makes the model pay more attention to minority classes and focus on pixels difficult to classify. The categoricalfocalcrossentropy loss function is implemented from the keras.losses library.

6.2.2. Model Design Choices

Input and output shapes: The task of multi-label segmentation and the nature of greyscale MRI images require the output mask shape of the models to have a size of (128,128, 4) and the input shape to have a size of (128,128, 1), as shown in Figures 8, 9, and 10 for UNet, Res-UNet and Attention Res-UNet respectively. This in turn reduces the total number of parameters in the model.

Activation function: The softmax classifier is used as the activation function in the output layer of all the models, as it is equipped to run classification/segmentation for multi-labeled prediction.

UNet number of parameters: 31401556

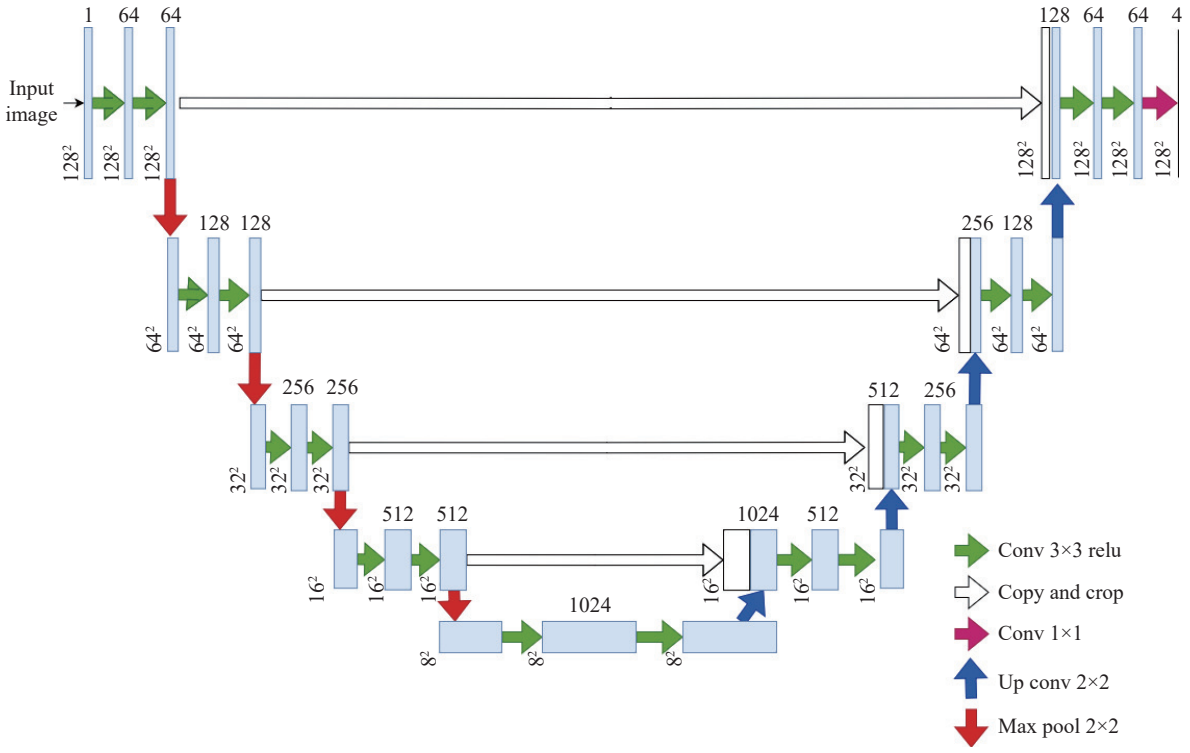

Figure 8. UNet Architecture.

Res-UNet number of parameters: 33157140

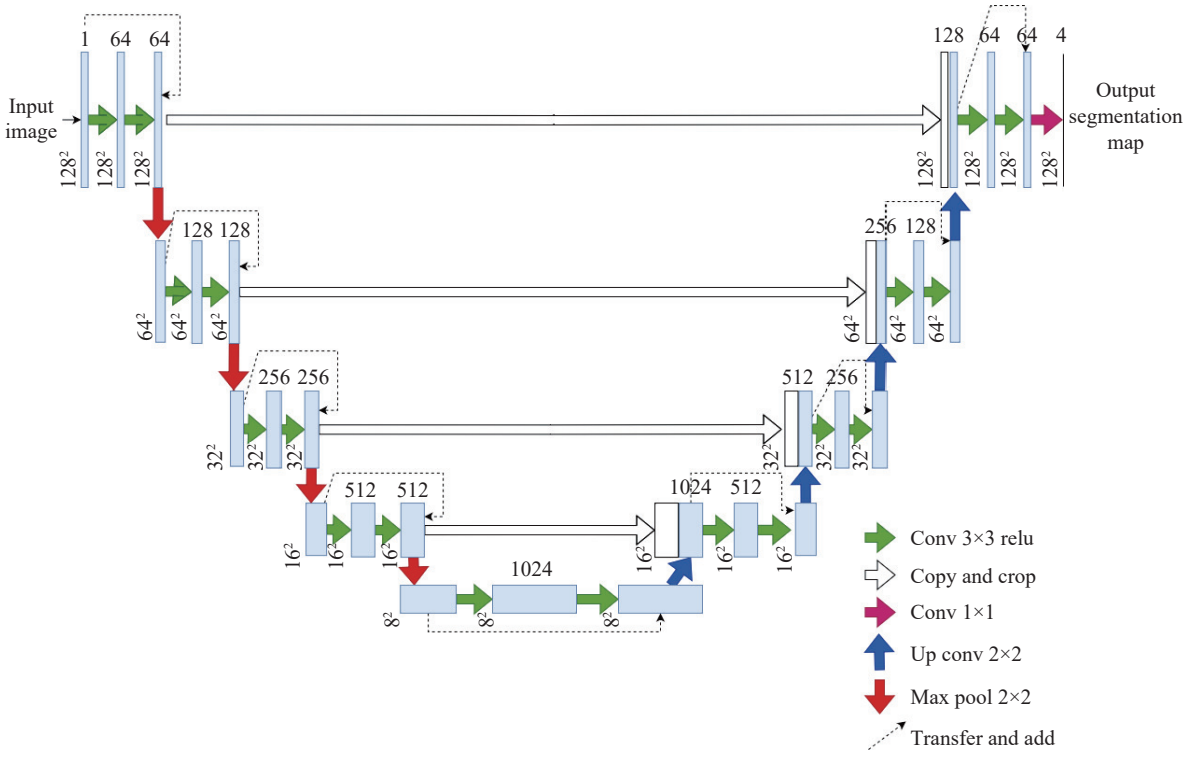

Figure 9. Res-UNet Architecture.

Attention Res-UNet number of parameters: 39089304

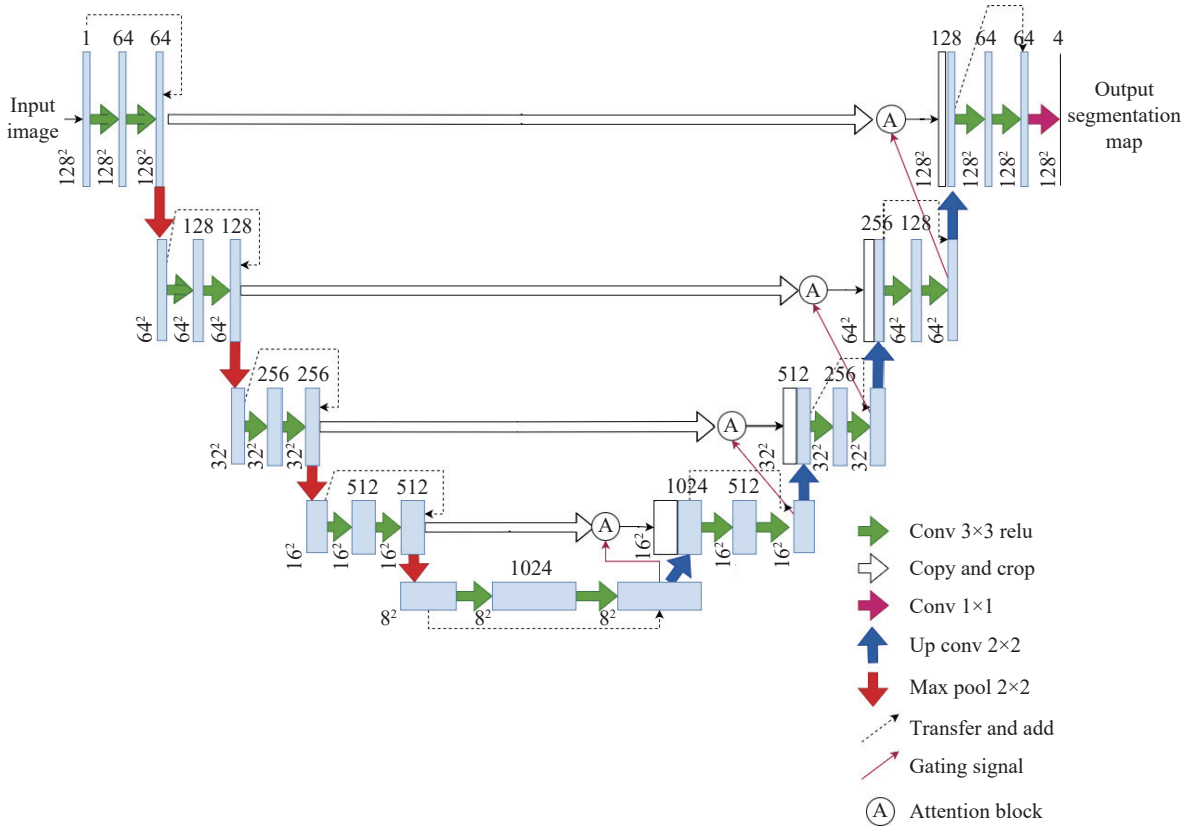

Figure 10. Attention Res-UNet Architecture.

6.2.3. Model Compiling and Fitting

All the models are compiled with the Adam optimizer at the initial learning rate of 1e-5, and are fitted with early stopping.

6.3. Results

Model training times and epochs are listed in Table 5. The UNet has the shortest training duration of 19 minutes, but requires 86 training epochs to reach convergence. In contrast, the Res-UNet has a longer training duration of 25 minutes and 20 seconds, and completes 98 training epochs before converging. The attention Res-UNet has the longest training duration of 27 minutes and 48 seconds, and reaches convergence after 83 training epochs.

Table 5 Execution time and epochs of trained models for Multi-label Heart Segmentation

Models	Execution Time	Epoch
UNet	19 min	86
Res-UNet	25 min 20 sec	98
Attention Res-UNet	27 min 48 sec	83

These results illustrate the trade-offs between the training time and the number of epochs required for these models. The UNet can be trained relatively quickly, but more epochs are required. The Res-UNet and attention Res-UNet take more training time, but require fewer epochs to achieve convergence.

Figure 11 shows the change in the categorical focal crossentropy over epochs for validation data of three models. All models converge similarly, starting with a high initial loss that rapidly decreases within the first 10 epochs. Afterward, the models exhibit noticeable fluctuations in the loss, and the Res-UNet shows fewer fluctuations compared to the others. Overall, convergence patterns of the models are similar. Table 6 provides the precision and recall values for each class predicted by the three models.

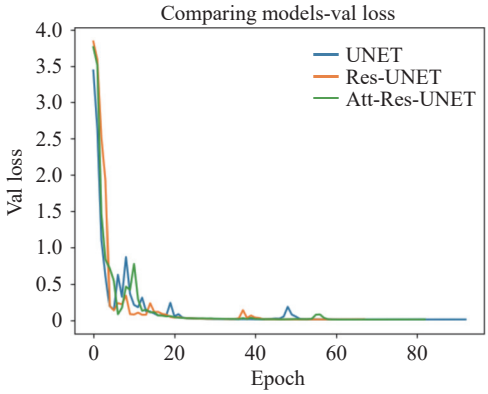

Figure 11. Convergence for trained models on Heart Segmentation

Table 6 Precision and Recall score for each class by three models

	Precision				Recall			
Models	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
UNet	0.99	0.91	0.89	**0.96**	0.99	0.91	**0.89**	0.94
Res-UNet	0.99	**0.92**	0.89	0.95	0.99	**0.92**	**0.89**	0.94
Attention Res-UNet	0.99	0.91	0.89	0.94	0.99	0.91	0.88	**0.95**

Class-wise performance evaluation

• **Class 0 (background):** All models achieve high precision scores, indicating that they are good at minimizing false positives for the background class. Note that the UNet and Res-UNet achieve the highest recall, suggesting that they capture most of the background pixels. The UNet achieves the highest dice coefficient and IoU, indicating its accuracy in identifying the background class.

• **Class 1 (RV cavity):** The Res-UNet achieves the highest precision for this class, indicating its accuracy in positive predictions. It also has the highest recall, meaning that most of the RV cavity pixels are captured. The Res-UNet has the highest dice coefficient, indicating accurate spatial predictions, while the UNet has the highest IoU.

• **Class 2 (myocardium):** The UNet has the highest precision for the myocardium, indicating accurate positive predictions. The Res-UNet has the highest recall and captures the most myocardium pixels. The UNet achieves the highest dice coefficient and IoU for the myocardium.

• **Class 3 (LV cavity):** The attention Res-UNet achieves the highest precision for the LV cavity, indicating its proficiency in minimizing false positives. It also has the highest recall, suggesting that most of the LV cavity pixels are captured. The UNet achieves the highest dice coefficient and IoU for the LV cavity.

Details of the class-wise performane for three models are shown in Table 7, measured by both Dice and IoU scores.

Table 7 Dice and IoU score for each class by three models

	Dice				IoU			
Models	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
UNet	0.993	0.906	**0.893**	**0.951**	**0.987**	0.829	**0.807**	**0.907**
Res-UNet	0.993	**0.92**	0.888	0.944	**0.987**	**0.852**	0.799	0.895
Attention Res-UNet	0.993	0.908	0.884	0.945	0.985	0.831	0.792	0.895

Accuracy and Loss

The overall accuracy and loss for the three models are presented in Table 8.

Table 8 Accuracy and Loss score by the three models

Models	**Accuracy**	**Loss**
UNet	**98.41%**	**1.00%**
Res-UNet	**98.41%**	1.09%
Attention Res-UNet	98.28%	1.44%

• The UNet achieves the highest accuracy of 98.41%, indicating its proficiency in overall pixel-level classification. The UNet also has the lowest loss of 1.00%, suggesting that the difference is minimized between predicted

and ground truth masks.

• The Res-UNet achieves the similar accuracy of 98.41%, but has a slightly higher loss of 1.09%.

• The attention Res-UNet has the accuracy of 98.28% and the highest loss of 1.44%.

In summary, the results show that each model excels in the following different aspects.

• The UNet demonstrates the high accuracy, low loss, and strong performance in capturing the background, myocardium, and LV cavity classes.

• The Res-UNet achieves high precision and recall for the RV cavity, and the highest dice coefficient for class 1 (RV cavity).

• The attention Res-UNet excels in precision and recall for the LV cavity and class 3 (LV cavity).

6.3.1. Discussion

Figure 12 shows four examples with the given image and ground truth mask followed by predictions from the three models.

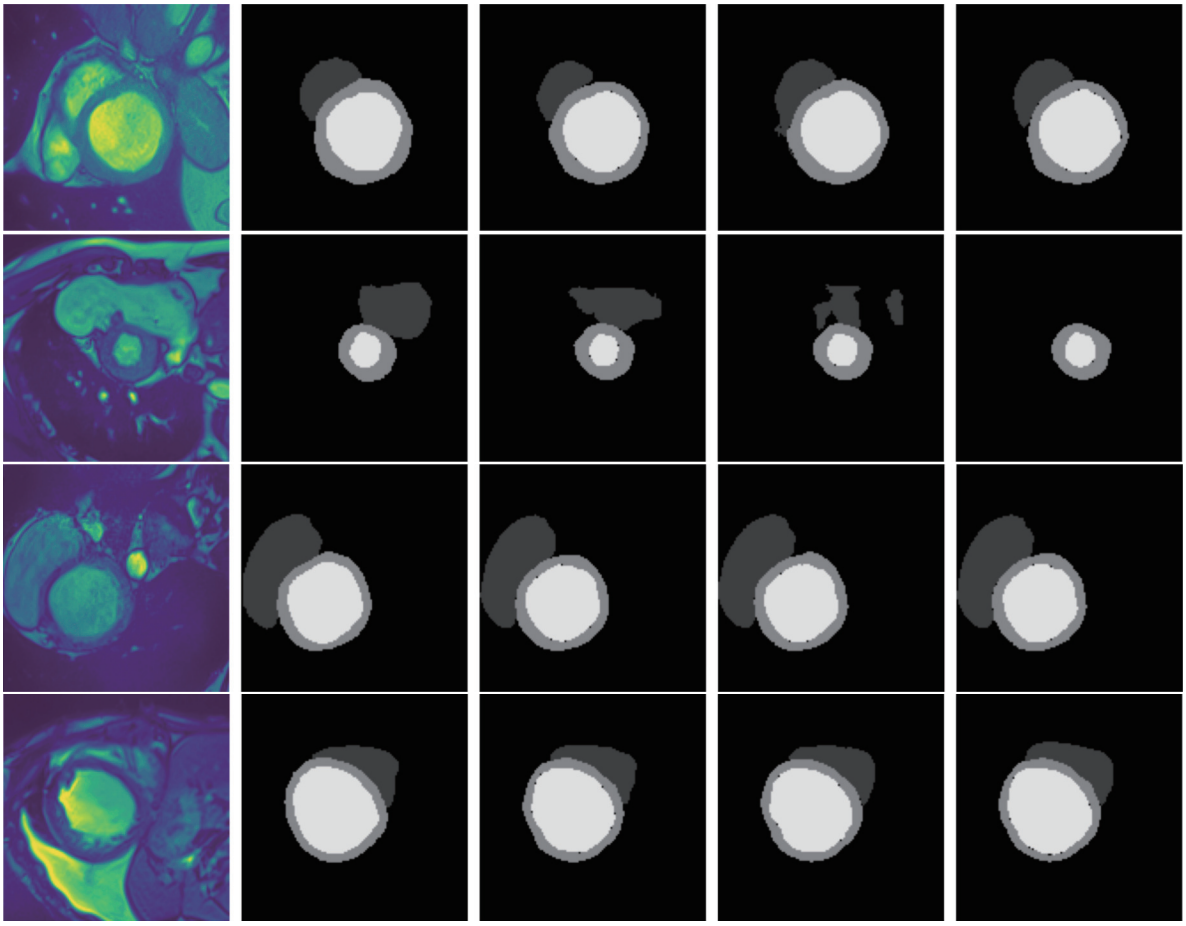

Figure 12. Segmentation results by the three models for four different examples, from left to right are the input images, ground-truth, segmentation results by UNet, Res-UNet and Attention Res-UNet.

The trained models (the UNet, Res-UNet, and attention Res-UNet) exhibit acceptable results in producing masks similar to the ground truth in the multi-class image segmentation task, involving the classes myocardium (Class 2), LV cavity (Class 3), and RV cavity.

All three models generally perform well in producing accurate image segmentation masks, particularly for the Myocardium (Class 2) and LV cavity (Class 3), due to the abundance of training examples and distinctive features. Note that all models struggle with the underrepresented RV cavity class (Class 1), resulting in frequent misclassifications, likely due to the limited training data of this class.

Despite these challenges, the UNet outperforms other models in capturing Class 1 pixels. This may be attributed to the UNet's lower focal loss score for Class 0, indicating that the problem of class imbalance is well handled and the focus is enhanced on the RV cavity class.

In summary, the models encounter typical challenges associated with imbalanced class distributions in multi-class image segmentation. The models excel with well-represented classes, but face difficulties with underrepre-

sented ones. The UNet shows promise in handling the imbalance problem, and techniques such as class balance and data augmentation could enhance performance across all classes.

7. Conclusions

In this paper, we have evaluated the performances of the UNet, Res-UNet and attention Res-UNet in solving three problems of the brain tumor, polyp and multi-label heart segmentation. All models have achieved acceptable segmentation results when compared to the ground-truth provided by the datasets. Differences are visible when the target masks become more complex in nature. The key findings of the study have been summarised as follows.

1) The UNet often misclassifies target classes as the background when overall target pixels are relatively small such as brain tumors or small polyp segments. The UNet also struggles with target segmentation when mask edge boundaries are intricate in nature. This points to the limitations of the UNet such as the vanishing gradient problem and the inability to put focus on hard-to-classify pixels.

2) The Res-Unet and attention Res-Unet are more suitable in handling complex and irregular structures, as both models are able to capture the complex boundaries in most cases. This is indicative of the residual connections introduced in the two models, and this mitigates the vanishing gradient problem.

3) The attention Res-Unet is more effective at tackling the problem of class imbalance, as it consistently achieves high recall values when solving all tasks. The model is able to predict more refined masks in most cases compared to the Res-UNet. Multi-label heart segmentation enforces these theories, as the mask images are less imbalanced compared to other cases, resulting in higher performances than the standard UNet model. The Res-Unet and attention Res-UNet perform similarly due to the exclusion of major classes under-representation. One of the three classes is often misclassified due to its scarcity in most of the images in the dataset. This indicates that datasets need to be more inclusive of all classes in order to make these robust models perform at their full potentials.

We have investigated three UNet architectures and applied them to solve three medical imaging problems. This is somewhat limited due to time constraints. In the future, more architectures will be explored, and more problems will be investigated. Other issues will also be investigated in future studies, such as the models' performance on images with different levels of noises and artifacts, the data augmentation techniques to improve the models' generalisation ability, and the feasibility and advantages of adopting transfer learning techniques [56, 57],

The implications of this work extend beyond the immediate research domain. This work sets a modern benchmark for segmentation techniques in the medical field, and offers future researchers valuable insights into the critical factors when applying the UNet, its variants and other deep learning methodologies to medical image analysis. To enhance this study, future work could focus on the application of the aforementioned models to three-dimensional medical images, as many medical datasets are inherently three-dimensional. Additionally, involving medical specialists in evaluating segmentation outputs could provide more refined and clinically relevant assessment. More loss functions and their effect on these models can be explored, thus adding to the reliability of the study and these models. Similar studies on more extensions of the UNet and the corresponding suitability can also be explored to enrich concrete guidelines.

Author Contributions: **Walid Ehab:** conception, data processing, implementation, result analysis and writing-up. **Lina Huang:** implementation, result analysis and writing-up. **Yongmin Li:** supervision, conception and writing-up. All authors have read and agreed to the published version of the manuscript.

Funding: No external funding was received for this work.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, *Munich*, *Germany*, *5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9351, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28
2. He, K.M.; Zhang, X.Y.; Ren, S.Q.; *et al*. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), *Las Vegas*, *NV*, *USA*, *27–30 June 2016*; IEEE: New York, 2016; pp. 770–778. doi: 10.1109/CVPR.2016.90
3. Maji, D.; Sigedar, P.; Singh, M.. Attention Res-UNet with Guided Decoder for semantic segmentation of brain tumors. Biomed. Signal Process. Control, **2022**, *71*: 103077. doi: 10.1016/j.bspc.2021.103077

4. Liu, L.L.; Cheng, J.H.; Quan, Q.; *et al*.. A survey on U-shaped networks in medical image segmentations. Neurocomputing, **2020**, *409*: 244−258. doi: 10.1016/j.neucom.2020.05.070
5. Haque, I.R.I.; Neubert, J.. Deep learning approaches to biomedical image segmentation. Inform. Med. Unlocked, **2020**, *18*: 100297. doi: 10.1016/j.imu.2020.100297
6. Lock, F.K.; Carrieri, D.. Factors affecting the UK junior doctor workforce retention crisis: An integrative review. BMJ Open, **2022**, *12*: e059397. doi: 10.1136/bmjopen-2021-059397
7. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; *et al*.. Active shape models-their training and application. Comput. Vis. Image Underst., **1995**, *61*: 38−59. doi: 10.1006/cviu.1995.1004
8. Litjens, G.; Kooi, T.; Bejnordi, B.E.; *et al*.. A survey on deep learning in medical image analysis. Med. Image Anal., **2017**, *42*: 60−88. doi: 10.1016/j.media.2017.07.005
9. Fraz, M.M.; Remagnino, P.; Hoppe, A.; *et al*.. Blood vessel segmentation methodologies in retinal images–a survey. Comput. Methods Programs Biomed., **2012**, *108*: 407−433. doi: 10.1016/j.cmpb.2012.03.009
10. Ren, S.Q.; He, K.M.; Girshick, R.; *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, *Montreal*, *Canada*, *7–12 December 2015*; MIT Press: Cambridge, 2015; pp. 91–99.
11. Valvano, G.; Santini, G.; Martini, N.; *et al*.. Convolutional neural networks for the segmentation of microcalcification in mammography imaging. J. Healthc. Eng., **2019**, *2019*: 9360941. doi: 10.1155/2019/9360941
12. Callaghan, M.F.; Josephs, O.; Herbst, M.; *et al*.. An evaluation of prospective motion correction (PMC) for high resolution quantitative MRI. Front. Neurosci., **2015**, *9*: 97. doi: 10.3389/fnins.2015.00097
13. Iglesias, J.E.; Liu, C.Y.; Thompson, P.M.; *et al*.. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging, **2011**, *30*: 1617−1634. doi: 10.1109/TMI.2011.2138152
14. Havaei, M.; Davy, A.; Warde-Farley, D.; *et al*.. Brain tumor segmentation with Deep Neural Networks. Med. Image Anal., **2017**, *35*: 18−31. doi: 10.1016/j.media.2016.05.004
15. Fritscher, K.D.; Peroni, M.; Zaffino, P.; *et al*.. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Med. Phys., **2014**, *41*: 051910. doi: 10.1118/1.4871623
16. Ma, J.L.; Wu, F.; Jiang, T.A.; *et al*.. Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images. Med. Phys., **2017**, *44*: 1678−1691. doi: 10.1002/mp.12134
17. Huda, W.; Slone, R.M. *Review of Radiologic Physics*, 2nd ed.; Lippincott Williams & Wilkins: Philadelphia, 2003.
18. Zaitsev, M.; Maclaren, J.; Herbst, M.. Motion artifacts in MRI: A complex problem with many partial solutions. J. Magn. Reson. Imaging, **2015**, *42*: 887−901. doi: 10.1002/jmri.24850
19. Plenge, E.; Poot, D.H.J.; Bernsen, M.; *et al*.. Super-resolution methods in MRI: Can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time?. Magn. Reson. Med., **2012**, *68*: 1983−1993. doi: 10.1002/mrm.24187
20. Rabbani, H.; Vafadust, M.; Abolmaesumi, P.; *et al*.. Speckle noise reduction of medical ultrasound images in complex wavelet domain using mixture priors. IEEE Trans. Biomed. Eng., **2008**, *55*: 2152−2160. doi: 10.1109/TBME.2008.923140
21. Wells, P.N.T.; Liang, H.D.. Medical ultrasound: Imaging of soft tissue strain and elasticity. J. Roy. Soc. Interface, **2011**, *8*: 1521−1549. doi: 10.1098/rsif.2011.0054
22. Duarte-Salazar, C.A.; Castro-Ospina, A.E.; Becerra, M.A.; *et al*.. Speckle noise reduction in ultrasound images for improving the metrological evaluation of biomedical applications: An overview. IEEE Access, **2020**, *8*: 15983−15999. doi: 10.1109/ACCESS.2020.2967178
23. Kamiyoshihara, M.; Otaki, A.; Nameki, T.; *et al*.. Congenital bronchial atresia treated with video-assisted thoracoscopic surgery; report of a case. Kyobu Geka, **2004**, *57*: 591−593
24. Aichinger, H.; Dierker, J.; Joite-Barfuß, S.; *et al*. *Radiation Exposure and Image Quality in X-Ray Diagnostic Radiology*: *Physical Principles and Clinical Applications*, 2nd ed.; Springer: Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-11241-6
25. Chen, J.Y.; Zheng, H.B.; Lin, X.; *et al*.. A novel image segmentation method based on fast density clustering algorithm. Eng. Appl. Artif. Intell., **2018**, *73*: 92−110. doi: 10.1016/j.engappai.2018.04.023
26. Sezgin, M.; Sankur, B.. Survey over image thresholding techniques and quantitative performance evaluation. J. Electron. Imaging, **2004**, *13*: 146−165. doi: 10.1117/1.1631315
27. Felzenszwalb, P.F.; Huttenlocher, D.P.. Efficient graph-based image segmentation. Int. J. Comput. Vis., **2004**, *59*: 167−181. doi: 10.1023/B:VISI.0000022288.19776.77
28. Beucher, S.. The watershed transformation applied to image segmentation. Scann. Microsc., **1992**, *1992*: 28
29. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall, Inc.: Englewood Cliffs, 1988.
30. Kass, M.; Witkin, A.; Terzopoulos, D.. Snakes: Active contour models. Int. J. Comput. Vis., **1988**, *1*: 321−331. doi: 10.1007/BF00133570
31. Kaba, D.; Salazar-Gonzalez, A.G.; Li, Y.M.; *et al*. Segmentation of retinal blood vessels using Gaussian mixture models and expectation maximisation. In *2nd International Conference on Health Information Science*, *London*, *UK*, *25–27 March 2013*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 105–112. doi: 10.1007/978-3-642-37899-7_9
32. Kaba, D.; Wang, C.; Li, Y.M.; *et al*.. Retinal blood vessels extraction using probabilistic modelling. Health Inf. Sci. Syst., **2014**, *2*: 2. doi: 10.1186/2047-2501-2-2
33. Kaba, D.; Wang, Y.X.; Wang, C.; *et al*.. Retina layer segmentation using kernel graph cuts and continuous max-flow. Opt. Express, **2015**, *23*: 7366−7384. doi: 10.1364/OE.23.007366
34. Dodo, B.I.; Li, Y.M.; Eltayef, K.; *et al*. Graph-cut segmentation of retinal layers from OCT images. In *11th International Joint Conference on Biomedical Engineering Systems and Technologies*, *Funchal*, *Madeira*, *Portugal*, *19–21 January 2018*; SciTePress: Setúbal, Portugal, 2018; pp. 35–42. doi: 10.5220/0006580600350042
35. Salazar-Gonzalez, A.; Kaba, D.; Li, Y.M.; *et al*.. Segmentation of the blood vessels and optic disk in retinal images. IEEE J. Biomed. Health Inform., **2014**, *18*: 1874−1886. doi: 10.1109/JBHI.2014.2302749
36. Salazar-Gonzalez, A.; Li, Y.M.; Kaba, D. MRF reconstruction of retinal images for the optic disc segmentation. In *1st International Conference on Health Information Science*, *Beijing*, *China*, *8–10 April 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 88–99. doi: 10.1007/978-3-642-29361-0_13
37. Salazar-Gonzalez, A.G.; Li, Y.M.; Liu, X.H. Optic disc segmentation by incorporating blood vessel compensation. In *IEEE Third*

International Workshop on Computational Intelligence in Medical Imaging, *Paris*, *France*, *11–15 April 2011*; IEEE: New York, 2011; pp. 1–8. doi: 10.1109/CIMI.2011.5952040
38. Dodo, B.I.; Li, Y.M.; Eltayef, K.; *et al*.. Automatic annotation of retinal layers in optical coherence tomography images. J. Med. Syst., **2019**, *43*: 336. doi: 10.1007/s10916-019-1452-9
39. Dodo, B.I.; Li, Y.M.; Kaba, D.; *et al*.. Retinal layer segmentation in optical coherence tomography images. IEEE Access, **2019**, *7*: 152388−152398. doi: 10.1109/ACCESS.2019.2947761
40. Wang, C.; Kaba, D.; Li, Y.M.. Level set segmentation of optic discs from retinal images. J. Med. Bioeng., **2015**, *4*: 213−220. doi: 10.12720/jomb.4.3.213-220
41. Wang, C.; Wang, Y.X.; Li, Y.M.. Automatic choroidal layer segmentation using Markov random field and level set method. IEEE J. Biomed. Health Inform., **2017**, *21*: 1694−1702. doi: 10.1109/JBHI.2017.2675382
42. Drozdzal, M.; Chartrand, G.; Vorontsov, E.; *et al*.. Learning normalized inputs for iterative estimation in medical image segmentation. Med. Image Anal., **2018**, *44*: 1−13. doi: 10.1016/j.media.2017.11.005
43. Li, H.; Zeng, N.Y.; Wu, P.S.; *et al*.. Cov-Net: A computer-aided diagnosis method for recognizing COVID-19 from chest X-ray images via machine vision. Expert Syst. Appl., **2022**, *207*: 118029. doi: 10.1016/j.eswa.2022.118029
44. Zeng, N.Y.; Li, H.; Peng, Y.H.. A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease. Neural Comput. Appl., **2023**, *35*: 11599−11610. doi: 10.1007/s00521-021-06149-6
45. Li, H.; Wu, P.S.; Zeng, N.Y.; *et al*.. A survey on parameter identification, state estimation and data analytics for lateral flow immunoassay: From systems science perspective. Int. J. Syst. Sci., **2022**, *53*: 3556−3576. doi: 10.1080/00207721.2022.2083262
46. Badrinarayanan, V.; Kendall, A.; Cipolla, R.. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., **2017**, *39*: 2481−2495. doi: 10.1109/TPAMI.2016.2644615
47. Kao, P.Y.; Ngo, T.; Zhang, A.; *et al*. Brain tumor segmentation and tractographic feature extraction from structural MR images for overall survival prediction. In *4th International Workshop on Brainlesion*: *Glioma*, *Multiple Sclerosis*, *Stroke and Traumatic Brain Injuries*, *Granada*, *Spain*, *16 September 2018*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 128–141. doi: 10.1007/978-3-030-11726-9_12
48. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; *et al*.. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods, **2021**, *18*: 203−211. doi: 10.1038/s41592-020-01008-z
49. McConnell, N.; Miron, A.; Wang, Z.D.; *et al*. Integrating residual, dense, and inception blocks into the nnUNet. In *IEEE 35th International Symposium on Computer Based Medical Systems*, *Shenzen*, *China*, *21–23 July 2022*; IEEE: New York, 2022; pp. 217–222. doi: 10.1109/CBMS55023.2022.00045
50. Ndipenoch, N.; Miron, A.; Wang, Z.D.; *et al*. Simultaneous segmentation of layers and fluids in retinal OCT images. In *15th International Congress on Image and Signal Processing*, *BioMedical Engineering and Informatics*, *Beijing*, *China*, *5–7 November 2022*; IEEE: New York, 2022; pp. 1–6. doi: 10.1109/CISP-BMEI56279.2022.9979957
51. Ndipenoch, N.; Miron, A.D.; Wang, Z.D.; *et al*. Retinal image segmentation with small datasets. In *16th International Joint Conference on Biomedical Engineering Systems and Technologies*, *Lisbon*, *Portugal*, *16–18 February 2023*; SciTePress: Setúbal, Portugal, 2023; pp. 129–137. doi: 10.5220/0011779200003414
52. Buda, M. Brain MRI segmentation.
53. Pedano, N.; Flanders, A.E.; Scarpace, L.; *et al*. The cancer genome atlas low grade glioma collection (TCGA-LGG) (version 3). The Cancer Imaging Archive, 2016.
54. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; *et al*.. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph., **2015**, *42*: 99−111. doi: 10.1016/j.compmedimag.2015.02.007
55. Bernard, O.; Lalande, A.; Zotti, C.; *et al*.. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?. IEEE Trans. Med. Imaging, **2018**, *37*: 2514−2525. doi: 10.1109/TMI.2018.2837502
56. Ma, G.J.; Wang, Z.D.; Liu, W.B.; *et al*.. A two-stage integrated method for early prediction of remaining useful life of lithium-ion batteries. Knowl.-Based Syst., **2023**, *259*: 110012. doi: 10.1016/j.knosys.2022.110012
57. Ma, G.J.; Wang, Z.D.; Liu, W.B.; *et al*.. Estimating the state of health for lithium-ion batteries: A particle swarm optimization-assisted deep domain adaptation approach. IEEE/CAA J. Autom. Sin., **2023**, *10*: 1530−1543. doi: 10.1109/JAS.2023.123531

Citation: Ehab, W.; Huang, L.; Li, Y. UNet and Variants for Medical Image Segmentation. *International Journal of Network Dynamics and Intelligence*. 2024, 3(2), 100009. doi: 10.53941/ijndi.2024.100009