

Kidney segmentation in MRI images using deep learning: A comparison between U-Net and Attention U-Net (June 2023)

Francisco Freiria, Rafael Rodrigues, Plinio Moreno, Antonio M. G. Pinheiro

Abstract—The increasing prevalence of Chronic Kidney Disease represents a significant public health challenge, affecting more than 10% of the population. Magnetic Resonance Imaging biomarkers have been shown to be sensitive to pathophysiological changes, are non-invasive, and may help reduce the need for biopsies and the risk of complications in patients with Chronic Kidney Disease. Total volume is the most assessed parameter in patients with Autosomal Dominant Polycystic Kidney Disease and helps monitor the progression of Chronic Kidney Disease. Kidney segmentation is an essential step for assessing renal volume. However, it often relies on manual segmentation, which is a time-consuming and highly subjective task. Deep learning methodologies have contributed to developing algorithms that provide accurate, inexpensive, and user-independent results. This dissertation explores deep learning-based approaches for kidney segmentation, notably U-Net and Attention U-Net. Both architectures were also tested with the standard cross-entropy loss function and a focal cross-entropy loss function. Finally, the proposed solutions were used to perform both 2-class and 3-class segmentation as an attempt to improve the segmentation of border regions. A dataset provided by the COST action PARENCHIMA was used for both model training and validation, with a leave-one-out cross-validation scheme. The best Dice coefficient obtained on the testing set was 0.966%.

Impact Statement—This paper studies the application of deep learning methodologies for kidney segmentation in MRI scans, addressing the public health challenge of Chronic Kidney Disease. The study focuses on total kidney volume measurement, a crucial parameter for disease progression monitoring. The research provides accurate and user-independent results by using automated deep learning-based solutions, thus eliminating the need for time-consuming manual segmentation. The reported findings may have a significant impact in clinical practice, offering automation and accuracy in kidney segmentation, therefore reducing the burden on healthcare professionals, minimizing complications from invasive procedures, and improving Chronic Kidney Disease treatment and follow-up. The obtained results support the integration of deep learning-based algorithms into clinical settings, for a more efficient and reliable analysis of kidney pathophysiology.

Index Terms—Chronic Kidney Disease, Deep Learning, Kidney Segmentation, Magnetic Resonance Imaging.

I. INTRODUCTION

THE Chronic Kidney Disease (CKD) is a growing public health challenge. Magnetic Resonance Imaging (MRI) biomarkers offer non-invasive assessments of typical CKD pathophysiological changes associated with inflammation, fibrosis, oxygenation, and microstructure [1]. They provide

valuable information for disease monitoring disease and help reduce the need for biopsies and other invasive measurements. The PARENCHIMA COST action aimed at making their clinical acceptance wider through the standardization and availability of kidney MRI biomarkers. This COST action provided the data used in this study.

Total kidney volume (TKV) measurement is the most assessed parameter in patients with Autosomal Dominant Polycystic Kidney Disease (ADPKD) [2] and helps monitor the progression of CKD, among others. Renal volumetry may be obtained through accurate kidney segmentation. Although manual segmentation is still the most common approach, it is a time-consuming and highly operator-dependent task. Hence, there is a need for reliable automated methods. Several problems must be addressed when segmenting the renal image. Among them is the similarity of the intensity of the renal tissues to the adjacent organs. Thus, intensity threshold-based methodologies often fail to provide a robust segmentation. Deep learning techniques have been increasingly applied to medical imaging tasks and are becoming the leading approach to solving segmentation problems in medical images. Deep learning approaches can potentially analyze an image and provide reproducible and robust segmentation and volume measurements, as they can learn important image features to perform pixel-wise classification. Volumetry measurements can be assessed through 2D or 3D images. Since the width resolution is greater than the spatial resolution, segmenting the kidney in 3D images is not justified, which takes more time and requires more computational power.

Currently, the most used approaches in computer-aided diagnosis and the automated segmentation of medical images are based on Convolutional Neural Networks (CNN). There are architectures composed of convolution layers capable of obtaining important features from a given image by reducing its complexity. An example of a CNN is the U-Net, which is the most used model in medical image analysis and segmentation tasks.

The biggest problem for a successful implementation is having large datasets. A particular benefit of this architecture is that it does not require a large dataset compared to other architectures. Another common problem in kidney segmentation comes from class imbalance. Automated kidney segmentation is still challenging since the kidney occupies a small fraction of the image (e.g. $< 3\%$). Adding an attention mechanism to focus learning on most regions can help overcome this problem. The literature also suggests implementing Coarse-

to-Fine segmentation approaches [3] to address the problem of class imbalance and improve the border region between classes.

This dissertation presents an approach using Deep Learning methods for kidney segmentation in abdominal MRI images. A dataset of 21 manually annotated images is used to obtain the ground truth. An automated approach using a modified U-Net and Attention U-Net for kidney targeting has been developed and tested, which may further help evaluate kidney diseases.

Eight segmentation algorithms were implemented and compared:

- U-Net with cross-entropy loss function for binary and multi-class kidney segmentation;
- U-Net with focal loss function for binary and multi-class kidney segmentation;
- Attention U-Net with cross-entropy loss function for binary and multi-class kidney segmentation;
- Attention U-Net with focal loss function for binary and multi-class kidney segmentation.

II. THEORETICAL CONTEXT

A. Image segmentation

Image segmentation is the process of dividing an image into multiple non-overlapping regions according to certain criteria. The division of the image into a certain region or set of pixels reduces the analysis area. It facilitates and optimizes the process of searching for relevant features according to the imposed criteria [4]. This process results in a set of image segments that cover the original image when rejoined. Relevant image features must be highlighted to make image analysis easier and more meaningful. The main goal of segmentation algorithms is to make a pixel-wise prediction.

Medical imaging segmentation in the context of deep learning is a task that faces several challenges, and the formulation of several hypotheses should be raised. A widespread problem in medical imaging is the lack of labeled training data [5].

B. Magnetic Resonance Imaging

MRI provides complete kidney function and anatomy information. It allows for precisely visualizing the kidney's state and its constituent parts, such as the cortex, medulla, and pelvis. It will enable the identification of renal lesions, tumors, and small masses but is unsuitable for identifying calcification, including stones [6].

A magnetic resonance image is captured using a magnetic field that aligns the individual's free water protons along the magnetic field axis. A radio frequency (RF) antenna is placed over the area that is supposed to capture the image, which releases energy pulses. These RF pulses align protons to the angle of the magnetic field, causing the protons to spin in phase with each other creating resonance. Milliseconds after the RF pulse burst, the nuclei return to resting alignment through various relaxation processes which release RF energy. MRI captures released energy to generate an image [7]. Fourier transform converts the frequency information given by the signal from each location on the image plane to the

corresponding intensity levels, displayed as shades of gray in a matrix of pixels.

Several images can be created by varying RF pulse sequences applied or collected. They can be changed by the repetition time (TR), which is the amount of time between successive pulse sequences applied on the same slice, and by the echo time (TE), which is the time between the delivery of the RF pulse and the reception of the echo signal [8].

Several types of MRI sequences exist. The most commonly used are T1-weighted and T2-weighted, varying in TR and echo time (TE). T1-weighted images are produced using short TE and TR instead of T2-weighted images, which are produced using long TE and TR times. T1, the longitudinal relaxation time, is the time constant that determines the rate at which the spinning protons (excited) return to equilibrium and realign with the external magnetic field. T2 (transverse relaxation time) is the time constant that determines the rate at which spinning (excited) protons reach equilibrium or go out of phase with each other, causing them to lose phase coherence between nuclei, spinning perpendicular to the main magnetic field [8]. T1 images usually highlight a type of adipose tissue within the body, while T2 images highlight adipose tissue and water within the body.

The most significant advantage of using MRI for disease control is to provide 3D kidney visualization with high spatial resolution without radiation. MRI creates images with high contrast between soft tissues.

C. From deep learning to U-Net models in kidney segmentation

In recent years, DL models created segmentation models with notable performance improvements, causing a paradigm shift in the medical image segmentation field [9].

It was known that convolutional operations have applications in many different computer vision tasks due to their high representational power, and from that point of view CNNs, since they are composed of convolutional layers, have become dominant in various computer vision tasks [10] resulting in being related to the state-of-the-art in medical image segmentation. U-Net is already the primary tool for medical image segmentation tasks [11].

U-Net is a U shaped architecture that uses descending convolutions to obtain spatial information to compose a map of low-level features, and ascending convolutions to use this map to get images back to their original size with detailed object boundaries.

D. U-Net

U-Net is a deep learning architecture designed to solve the lack of data problems inherent in medical image segmentation. It is, therefore, an architecture designed to perform a pixel-wise classification.

In addition to the traditional methods, several CNN architectures were studied. Still, due to the lack of training data, they did not answer the evolution of deep learning methods in the medical imaging context.

The U-Net architecture is divided into two paths, the first being called the down-sampling path. The down-sampling path is the contraction path, also called the encoder, composed of convolution and pooling layers. This type of operation allows extracting high-level feature structures from the input data. As you advance along the architecture, the size of the images is compressed while the depth increases. This makes it possible to increase the receptive field, allowing the filters to capture larger areas since the max-pooling operations remove less important pixels, but the spatial information decreases. Thus, the encoder generates feature maps which are low-resolution representations of the input image [12].

The second path of this architecture is the up-sampling path, also called the decoder. This converts low-resolution images into high-resolution images, representing the original image's pixel-by-pixel segmentation in the output layer. The primary operations of the decoder are the transposed convolution operations, which are up-sampling operations with learnable parameters. The learnable parameters come from the kernel matrix. They are adjusted throughout the training process. Each image pixel produced in the down-sample path is multiplied by the kernel matrix, generating a matrix of a higher order for each pixel. Finally, all matrices are summed, obtaining the output of the transposed convolution.

Along the up-sampling path, in each layer, the width and length of the image are doubled while the number of channels is halved.

Another characteristic of this U-shaped architecture is that spatial information from the down-sample path is concatenated into the up-sampling path. This feature introduces an improvement in spatial location, represented by the grey arrows in figure 1, representing the schematic representation of the originally proposed U-Net architecture.

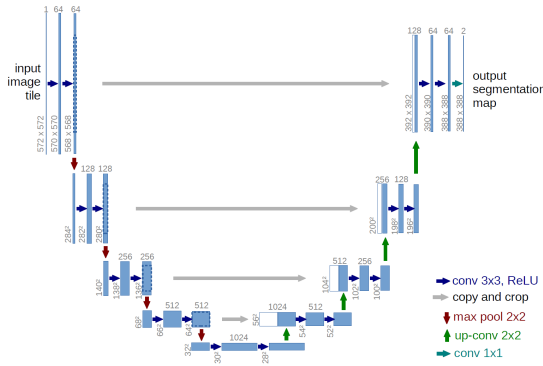


Fig. 1. Schematic representation of the original U-Net architecture. Image retrieved from [13].

E. Attention U-Net

The Attention U-Net model is an architecture that applies attention methods to the U-Net architecture described in the previous subsection. Attention mechanisms may highlight only the relevant actions during training, reducing the computational resources wasted on irrelevant activation and providing better network generalization.

To have a vision of the motivation and meaning of the implementation of this architecture in the segmentation of biomedical image context, the concept of this attention method will be indexed.

Human biological systems inspire attention mechanisms as they focus on learning relevant parts of the process when processing large amounts of information. Attention means focusing on what we care about and disregarding what is less important.

Although these mechanisms can be classified according to 4 different criteria [14], their implementation in this model concerns the smoothness of attention.

Thus, hard attention is a type of attention mechanism that highlights the relevant part of the image by cropping the relevant part. In other words, the training is done with cropped images of relevant parts of the original image or not so that the training can understand which parts to pay attention to. This implies that part of the training is not differentiable when the model is given part of the non-relevant image, forcing it to resort to reinforcement learning algorithms. Consequently, backpropagation mechanisms could not be used.

Soft attention is a type of attention that can be done in the training process itself, as opposed to hard attention. In the case of image segmentation, it assigns greater relevance throughout the training process to the relevant parts of the image, giving them greater weight and less weight to less relevant parts, giving them less weight. In other words, the weights are also trained to pay more attention to the relevant parts of the image. Since this process occurs throughout the training process, backpropagation mechanisms can be used.

Attention U-Net uses a soft attention mechanism integrated into the U-Net architecture. They are integrated after skip connections along the up-sampling path, so the encoder has the same architecture as the original U-Net.

The assignment of weights by this mechanism results from the concatenation of two inputs, one that provides spatial information (x) that gives context and that comes from the down-sampling path (skip connections), with the feature map (g) that provides feature representation as rounded parts, sharp edges, texture, etc., coming from the previous layer (deeper layers). So, Attention blocks give up-convolution layers the best of two worlds: The spatial information from the down-sampling path and initial layers (Skip connection) and the feature representation from deeper parts of the network.

Figure 2 illustrates the attention gate used by the attention U-Net model.

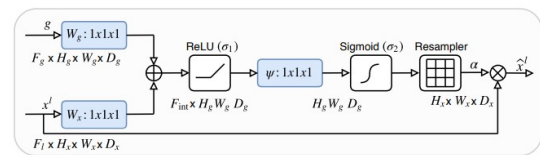


Fig. 2. Schematic representation of Attention mechanism architecture. Image retrieved from [15].

Since the inputs do not have the exact dimensions, because the g signal always comes from a deeper layer than x , resizing them to apply the arithmetic sum identified in the figure

above is necessary, the resizing is done through W_g and W_x convolution operations. In W_g , the convolution is done with a stride of (1,1) to keep the image dimensions and double the number of filters. In W_x , the convolution is done with a stride of (2,2) to reduce the image size to half and maintain the number of filters.

With the sum operation, the aligned weights get much larger assigning higher weights to more relevant pixels.

The ReLU activation function is applied to the sum result, in which the values of weights ≤ 0 take the value of 0, and for weights greater than zero, their maximum value. Then a convolution with only one filter is applied to obtain a dimension tensor $(h, w, 1)$, and its values are the new weights assigned to each image pixel.

Since these weights are linearly mapped in R , it is necessary to normalize these values to the scale of $[0, 1]$ to represent a probability. Therefore, the sigmoid function is applied.

The final result of the attention mechanism comes from the multiplication of this signal produced by the sigmoid activation function by the input x , both must have the same size. To do so, it will still be necessary to resample the signal produced by the sigmoid, which is done by an up-sampling operation. Multiplication by vector x scales the weights based on their relevance.

F. Metrics

a) *Confusion matrix*: The confusion matrix, although not a metric, is a table that summarizes the results of the classification predictions and produces the parameters used to calculate the evaluation metrics. This matrix shows how many predictions are correct or incorrect per class, allowing you to understand which classes are confused with which other classes. Each column of the matrix represents instances in the actual class, and each row represents the instances in the predicted class or vice versa. Figure 3 shows an example of a binary confusion matrix.

		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

Fig. 3. Visual representation of confusion matrix. Image retrieved from [16].

In the case of binary classification, the confusion matrix is a 2×2 matrix where **TP** denotes true positives (predicted correctly as positive), **TN** denotes true negatives (predicted correctly as negative), **FP** denotes false positives (mispredicted as positive) and **FN** denotes false negatives (mispredicted as negative) in a sense that predicted values are described as positives or negatives and actual values described as true or false.

b) *Accuracy*: In image segmentation, accuracy is the metric that informs the percentage of correct pixels between the model output for a given image and the corresponding ground truth. The equation 1 corresponds to the accuracy calculation equation.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

This is not a very efficient metric for kidney segmentation due to the class imbalance problem. However, evaluating the model's performance is an important metric since high values are already expected.

c) *Jaccard index*: Jaccard index or Intersection Over Union (IoU) urges the necessity of having a metric that quantifies the results regarding spatial similarity. Provides information on how well the algorithm could identify the true region of interest (ROI), measuring the overlap between the segmented image and the ground truth.

IoU is the most commonly used for comparing the similarity between two arbitrary shapes [17]. Encodes the shape properties on the same region of two images and then calculates a normalized measure focusing on their similarities. It is to be noted that this metric is even more relevant in multi-classification problems [18]. However, no strong correlation exists between minimizing the losses and an IoU improvement.

Equation 2 present the formula for the Jaccard index, and figure 4 gives the visual representation of IoU.

$$JaccardIndex = \frac{TP}{TP + FP + FN} \quad (2)$$

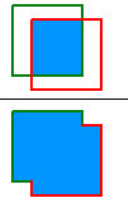
$$IoU = \frac{\text{area of overlap}}{\text{area of union}} = \frac{\text{area of overlap}}{\text{area of union}}$$


Fig. 4. Visual representation Intersection Over Union metric. Image retrieved from [19].

d) *Dice coefficient*: Dice coefficient (DSC) measures the overlap between the segmented image and the ground truth, thus taking into account the spatial arrangement, and therefore is a metric that resembles the Jaccard index.

The differences between the two metrics are that DSC tends to penalize errors closer to the average, and the IoU tends to penalize them closer to the worst case possible. DSC tends to emphasize the smaller regions of overlap.

It is also an important metric to assess the model's reproducibility and whether the results obtained in the training process are fighting the imbalance between the foreground and the background [20].

Equation 3 is represented the DSC equation, and figure 5 is the visual illustration.

$$DCS = \frac{2TP}{2TP + FN + FP} \quad (3)$$

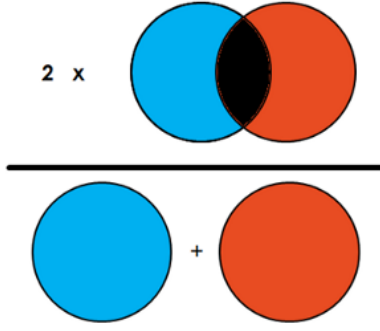


Fig. 5. Visual representation Dice coefficient metric. Image retrieved from [21].

e) *Recall*: Recall is the metric that presents the ratio between correctly classified pixels and all pixels that belong to the object of interest, detecting false positives. Therefore, useful to understand how well the algorithm explicitly identifies the ROI. In a medical context, this metric is important to avoid unnecessary treatment.

The equation that defines the recall using the elements of the confusion matrix is presented in 4.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

f) *Precision*: Precision is the metric that presents the ratio between true positives and all positives; it is, therefore, useful to understand how well the algorithm correctly identifies the pixels that belong to the object of interest, which in a medical context is important so that the evaluation of the segmented object can help in the diagnosis and medical evaluation.

The equation that defines the precision using the elements of the confusion matrix is presented in 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

All the presented metrics are commonly used in image segmentation, and each provides a different understanding of the quality of the segmentation.

The Jaccard index and the Dice coefficient provide a better understanding of the assessment because they measure the overlap between the segmented image and the ground truth, considering the spatial arrangement, size, and shape of the segmented object [22]. They can also help to identify regions where the algorithm has difficulties segmenting.

Accuracy does not consider the spatial arrangement of pixels, which in the case of class imbalance images, may not provide information of relevant interest. However, it is an excellent metric that helps monitor the evolution of the training process, allowing for algorithm improvement over time since high values are expected.

Recall and precision are more spatially aware than accuracy. Still, they do not consider the size and shape of the segmented object, which is very important in medical image segmentation. They are interested in being used as metrics because, in the case of precision, it allows the detection of false positives to be evaluated, which in a medical context

can lead to unnecessary treatment. Recall is the metric that best identifies whether the pixels of the object of interest is detected, which is important for medical evaluation in a medical context. Both metrics are based on the understanding and measure of relevance, hence are helpful in the quality performance description of the segmentation techniques [23].

G. Loss function

The loss function quantifies the difference between the expected result and the result produced by the model. It is one of the most critical parameters in neural network models. Any machine learning problem aims to minimize the loss function since it calculates the gradient to minimize the cost function of the loss function to update the weights and bias values through backpropagation algorithms. The adjustment of the hyperparameters has the purpose of reducing the loss value. This subsection discusses the loss functions used in this study.

a) *Cross Entropy*: In machine learning, cross-entropy (CE) defines a loss function that compares the predicted probability of a given pixel belonging to a given class with the class it belongs to and scores by penalizing the probabilities based on the expected value. This statistical distribution of labels plays an important role in evaluating training accuracy.

Thus, the CE loss function gives a measure of dissimilarity between the true probability and the probability estimated by the model, which is given by the following equation:

$$L_{CE}(y, \hat{y}) = - \sum_{c=1}^c \sum_{i=1}^N (y_{i,c} \log \hat{y}_{i,c}) \quad (6)$$

The predicted probability for each pixel belonging to a determined class ($\hat{y}_{i,c}$) is given by the soft-max equation, where $z_i(x)$ denotes the activation function in feature channel i at pixel position $x \in \Omega$ with $\Omega \subset \mathbb{Z}^2$, then penalizes it at each position the deviation of true value ($\hat{y}_{i,c}$) using:

$$E = \sum_{x \in \Omega} w(x) \log(\hat{y}_{i,c}(x)) \quad (7)$$

Where $C : \Omega \rightarrow \{1, \dots, K\}$ is the true label of each pixel and $w : \Omega \rightarrow \mathbb{R}$ is a weight map that gives to some pixels more importance in training [13]. The soft-max function calculates each pixel's energy function. Combined with the CE loss function, we obtain the approximate maximum function of the probability of a pixel belonging to the correct class.

In short, categorical cross-entropy is the loss function that uses the soft-max activation function to choose the one that produces the least loss among several classes.

b) *Focal loss function*: Focal loss (FL) deals with the front-background imbalance problems in object detection, making it very useful in problems where the background class is imbalanced compared to the object class. This loss function is similar to the CE loss function. Still, a modeling term is applied to focus learning on hard-to-classify examples, decreasing the weight on easy-to-classify examples and increasing the weights on hard-to-classify examples, depending on the training process confidence of predictions. Consequently, it improves performance on imbalanced datasets, as in kidney segmentation problems.

The modeling term is dynamically dimensioned, where the factor decays to zero as confidence in the correct class increases. This modeling term $(1-p)^\gamma$ is added to the standard cross-entropy criteria. Setting $(\gamma > 0)$ reduces the relative loss for well-classified examples $(p > 0.5)$ and increases focus on misclassified examples. γ is the focus parameter. As the CE loss function, the focal loss uses the soft-max equation to obtain the estimated probability (p) , which attributes great numerical stability [24].

The following equation gives the focal loss:

$$L_{Focal}(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (8)$$

III. DATASET AND METHODOLOGY

A. Dataset description

MRI images were taken from a database made available in the COST action CA16103 – “PARENCHIMA – Magnetic Resonance Imaging Biomarkers for Chronic Kidney Disease”, which aimed at obtaining renal volumetry from 2D MRI images. Assessing volumetry in 3D images is not justified since the longitudinal resolution of the kidneys is lower than the spatial resolution. The number of parameters associated with the training process of 3D image segmentation is much greater, which requires more time and greater computational power.

The dataset consists of 21 abdominal MRI images (.dcm files) and 21 ground truths obtained by manual kidney segmentation (.tif files) of each MRI image. All images were obtained from one patient.

MRI images were taken using the T1 VIBE acquisition methodology. Image acquisition through T1 means that magnetization has the same direction as the static magnetic field. They are taken with a Volumetric Interpolated Breath-hold Examination (VIBE) sequence that allows dynamic and high-resolution images in 30 seconds of apnea to minimize motion artifacts caused by respiratory movement but with high intrinsic contrast resolution for soft tissues. It has the advantage of improving the resolution of the Z-axis, which makes it possible to obtain high-quality multiplanar images and 3D reconstruction [25]. The VIBE acquisition methodology is a form of volumetric imaging using rapid 3D gradient-echo sequences [26]. All images have a slice thickness of 3.5 mm, a repetition time between 4.92 ms and 7.29 ms, and an echo time of 2.38 ms.

Specialists in the clinical environment captured the images, and the COST action verified the quality. The original size of 11 of 21 images is 320x260, and the remaining 10 with an original dimension of 320x240.

Given the limitations inherent to the size of the dataset, data augmentation techniques were used so that the training process would produce results of scientific relevance, introducing greater robustness.

Ground truth was annotated with two classes: the kidney, manually annotated, and the background. For the coarse-to-fine approach, ground truth has three classes: the kidney class, manually annotated; the contour class, obtained from the manual annotation of the kidney class and the background class with the imbalanced distribution.

B. Data Augmentation

Data augmentation techniques are recurrent and recommended by the literature on medical image segmentation problems. Inversions, rotations, translations, and variations in the grayscale were applied to the original images to create new data as suggested by the literature [13], [27]. The 20 images intended for the training process were transformed into a dataset composed of 1020 images.

The transformations and parameterizations applied to the original images are shown in table I.

TABLE I
DATA AUGMENTATION TRANSFORMATIONS.

Transformations		Parameters
Geometric	Inversion	Horizontal
	Rotation	[-5°, 5°]
	Translation	X axis [-7%, 7%]
		Y axis [-3%, 3%]
Gray Scale level	Brightness	[-10%, 40%]
	Contrast	[-10%, 40%]

C. Training

This section discusses the U-Net and Attention U-Net training model configurations. The hyperparameters and evaluation metrics used will be clarified.

The training of the models was carried out during 75 epochs with a batch size of 8. Empirical experiments chose the number of epochs, as the literature suggests [28], [13], [29], [30], to guarantee the convergence of the loss function without ever reaching overfitting. The training process lasted 75 epochs to balance managing computational resources and time with guaranteeing model convergence. The batch size was chosen to prevent memory problems.

The Adam optimizer was used because image segmentation is a complex computer vision task and requires a large dataset. The literature suggests that this optimizes large datasets and complex models compared to stochastic gradient descent with momentum (SGDM) [31].

It is a stochastic gradient descent optimization algorithm and dynamically adjusts timing parameters based on gradient descent updates. Considering the loss functions covered by the study, where pixel-wise differences between predicted and actual segmentation are calculated, adjusting the momentum parameters based on gradient descent updates helps minimize loss functions. It improves the convergence rate even with noisy data with sparse gradients.

The loss functions used in this study were Cross-Entropy and Focal loss. The choice of the loss function is a crucial point in deep learning architectures as they determine how the model is optimized throughout the training process and measure its performance. They have implications for the model's ability to generalize.

Data augmentation introduces robustness and new data tolerance to the model learning process by ensuring generalization ability.

A dropout rate was established as a regularization technique, reducing the interdependence of neurons to prevent memorization, standardization, and overfitting. It also makes the model more robust in the feature representation and removes the prominence between the training and test results.

Batch normalization was also used as a regularization technique, which helps to mitigate the displacement of internal covariance, adapting the model to different input distributions [32].

The initial learning rate is 10^{-2} ; we opted for an approach to the decay of the learning rate by step decay, in which it decays at a rate of 0.1 every 15 epochs, an approach that allowed us to avoid the model getting stuck at local minima, improve convergence by accelerating the training process and allowing unbiased training stability [33].

The table II summarizes the hyperparameters used in the proposed implementation.

TABLE II
HYPERPARAMETERS USED IN THE PROPOSED IMPLEMENTATION.

Model Parameters	U-Net	Attention U-Net
Starting LR	10^{-2}	
LR decay	drop = 0.1; Step decay = 15 epochs	
Optimization Strategy	Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-7}$)	
Epochs	75	
Batch size	8	
Dropout rate	0.1	
Batch normalization	True	

D. Test cross-validation

To demonstrate the validity and robustness of the implemented approaches and considering that the dataset is quite limited, a leave-one-experiment-out 10-k cross-validation was applied. This validation consists of leaving one dataset image for testing and the remaining 20 images for training. As such, 21 different training processes were performed. This avoids the bias introduced when testing a model with images with a high degree of similarity [34].

Cross-validation was performed ten times for each training set, totaling 210 training processes. This is a very useful validation technique when we do not have a large data size, as happens in this study [35].

IV. RESULTS

The segmentation performance for each model is given by evaluating the metrics obtained. Tables III and IV present the results for the binary and multi-class segmentation, respectively, of Dice coefficient (DSC) and Jaccard index (IoU).

The U-Net model using cross-entropy as a loss function achieved the highest performance at the Dice coefficient of 0.96577 ± 0.00871 and 0.93393 ± 0.01608 for the Jaccard Index.

Next, the Attention U-Net model presents a Dice coefficient of 0.96557 ± 0.00740 and a Jaccard index of 0.93354 ± 0.01377 . Preliminary results suggest that using the CE loss function improved the results compared to the results from the models in which the FL function was used. Similarly, we can conclude that the Attention U-Net results are the most significant.

A. Binary classification results

When looking at the experimental results for binary classification results from table III, we can conclude:

- Loss function: Both architectures achieved better segmentation results using the CE loss function. Models trained with FL converged and reached a local minimum too fast (during the first ten epochs), partially neglecting the rest of the training process. Figure 6 shows the graphs of the evolution of the loss function throughout the training process for the case in which the best segmentation performance is achieved in each model.
- Architecture: U-Net achieved higher segmentation results for both loss functions.
- Metrics: The architecture that presents the highest absolute percentage values is U-Net with a cross-entropy loss function. Observing the results obtained between U-Net and Attention U-Net, both using CE as a loss function, we conclude that U-Net obtained Accuracy 0.004% greater, Precision 0.574% greater, Dice coefficient 0.02% higher, Jaccard Index 0.04% higher and that the recall was higher 0.536% in Attention U-Net. Objectively, it is concluded that U-Net produced fewer false positives, given its greater precision, and Attention U-Net fewer false negatives, given its greater recall, since the remaining metrics present objectively equal results;
- Standard deviation: Attention U-Net models show lower standard deviation. Models that use the CE loss function also show lower standard deviation;
- Training metrics result: The Attention U-Net model with CE has the highest absolute percentage values of metrics during the training process;
- Best Results: Within the four models used to segment the kidney, evaluating the results of all metrics in the classification process, it is concluded that the model that presents the best results is the Attention U-Net model using CE as a loss function.

B. Three classification results

When looking at the experimental results for multi-class classification results from table IV, we can conclude:

- Loss function: Both architectures achieved better segmentation results using the CE loss function. Models trained with FL converged and reached a local minimum too fast (during the first ten epochs), partially neglecting the rest of the training process. Figure 6 shows the graphs of the evolution of the loss function throughout the training process for the case in which the best segmentation performance is achieved in each model.
- Architecture: Attention U-Net achieved higher segmentation results for both loss functions.

TABLE III

EVALUATION RESULTS OF DICE COEFFICIENT, JACCARD INDEX, ACCURACY, PRECISION, RECALL, LOSS FUNCTION AND CLASSIFICATION TIME OF THE TWO CLASSES SEGMENTATION MODELS.

Segmentation Model	<i>DCS</i>	<i>IoU</i>	Accuracy	Precision	Recall
U-NET + CE	0.966±0.009	0.934±0.016	0.997±0.001	0.967±0.012	0.965±0.019
U-NET + FL	0.960±0.011	0.923±0.020	0.997±0.001	0.980±0.010	0.941±0.023
ATT U-NET + CE	0.966±0.007	0.934±0.014	0.997±0.001	0.961±0.013	0.970±0.016
ATT U-NET + FL	0.944±0.094	0.902±0.094	0.996±0.004	0.970±0.096	0.920±0.095

TABLE IV

EVALUATION RESULTS OF DICE COEFFICIENT, JACCARD INDEX, ACCURACY, PRECISION, RECALL, LOSS FUNCTION AND CLASSIFICATION TIME OF THE MULTI-CLASS CLASSES SEGMENTATION MODELS

Segmentation Model	<i>DCS</i>	<i>IoU</i>	Accuracy	Precision	Recall
U-NET + CE	0.937±0.015	0.888±0.024	0.995±0.002	0.936±0.019	0.956±0.020
U-NET + FL	0.936±0.016	0.887±0.024	0.995±0.002	0.935±0.020	0.938±0.014
ATT U-NET + CE	0.939±0.013	0.891±0.020	0.995±0.002	0.939±0.017	0.939±0.013
ATT U-NET + FL	0.937±0.015	0.889±0.024	0.995±0.002	0.937±0.019	0.939±0.015

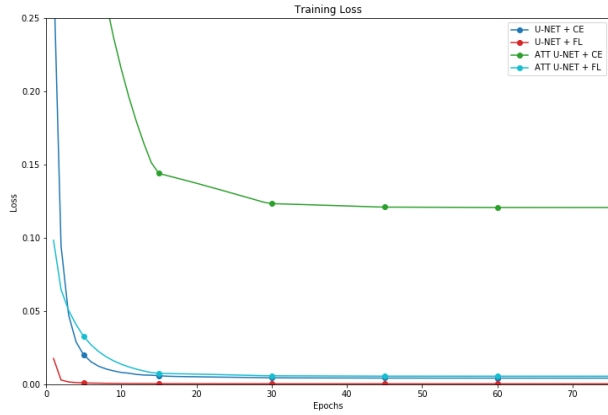


Fig. 6. Loss function evolution from the highest dice coefficient result for each of two class methods.

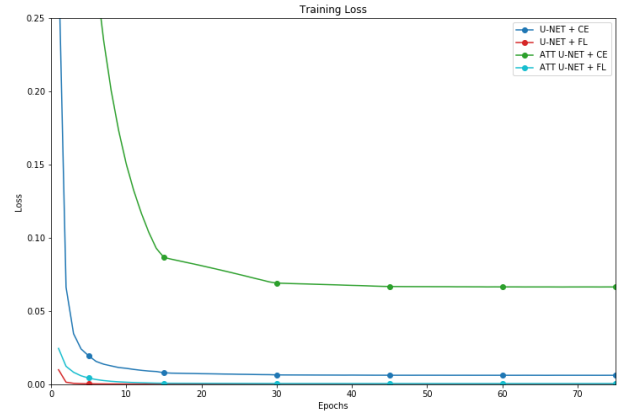


Fig. 7. Loss function evolution from the highest dice coefficient result for each of three class methods.

- Standard deviation: Attention U-Net models show lower standard deviation. Models that use the CE loss function also show lower standard deviation;
- Training metrics result: The U-Net model with CE has the highest absolute percentage values of metrics during the training process;
- Best Results: Within the four models used to segment the kidney, evaluating the results of all metrics in the classification process, it is concluded that the Attention U-Net model using CE as a loss function achieved better segmentation results.

C. Conclusion

- The highest achieved result for binary segmentation was produced by Attention U-Net with cross-entropy loss function;
- The highest achieved result for three class segmentation was produced by Attention U-Net with cross-entropy loss function;
- Binary segmentation models outperformed the proposed coarse-to-fine approach;

- Focal loss function best fits in multi-class segmentation approaches;
- Attention U-Net has shown more importance in multi-class segmentation approaches.

V. CONCLUSIONS

A. Achievements

In conclusion, this paper addressed the segmentation of kidneys in abdominal MRI images using the U-Net and Attention U-Net architectures. The objectives were successfully achieved, including obtaining an approach for kidney segmentation, comparing the results of U-Net and Attention U-Net, evaluating different loss functions, and exploring binary and multi-class segmentation as a coarse-to-fine approach.

The results demonstrated that the Attention U-Net model outperformed U-Net, contributing to state-of-the-art kidney segmentation. The Cross-Entropy loss function showed the best results among the studied functions, while the Focal Loss function requires further parameterization exploration. Binary segmentation yielded better results for small datasets with

limited epochs, while multi-class segmentation with a contour class did not significantly improve kidney segmentation.

Importantly, this study showed that a large dataset is not mandatory, and training neural networks with excessive epochs is not essential. These findings challenge some prevailing assumptions in the field.

The results support the consideration of Attention U-Net as a valuable architecture for developing methodologies and tools that utilize kidney segmentation in abdominal MRI images as a biomarker for kidney assessment. Further research can explore the optimization of the architecture's parameterization and investigate the potential of the Focal Loss function with refined parameters.

B. Future Works

Several avenues for future research can be explored for the proposed models to enhance their performance. This section outlines the key areas for the investigation to further advance kidney segmentation in MRI images using deep learning.

A. Optimization of the Proposed Algorithm

The algorithm that combines Attention U-Net with the Cross-Entropy loss function has demonstrated the best results in our study. However, further optimization can be performed to minimize the loss function. Specifically, studies on learning rate decay and the number of epochs should be prioritized to improve convergence and enhance the overall accuracy of the models.

B. Parameterization of the Focal Loss Function

Our research utilized the focal loss function with $\alpha = 0.25$ and $\gamma = 2$. However, further investigations should be conducted to explore different parameterizations of this loss function for medical image segmentation. Specifically, varying the alpha value while maintaining the gamma value constant can provide insights into the importance attributed to different classes.

C. Evaluation on Larger Datasets

While our study achieved promising results with a small dataset, evaluating the proposed models on larger datasets is imperative. Emphasizing datasets with a substantial number of samples will enable a quantitative analysis of the impact of dataset size on model performance. This investigation can shed light on the scalability and generalizability of the proposed methods.

D. Calculation of renal volumetry

To enhance the clinical relevance of our research, the calculation of the renal volumetry based on the obtained segmentation should be implemented. Performing instance segmentation on each kidney can yield accurate volume measurements, thereby contributing to potential clinical applications.

E. Exploration of Pre-trained Backbone Models

To leverage pre-existing knowledge learned from large-scale datasets, we suggest employing pre-trained weights of backbone models, such as ResNET, EfficientNet, or ImageNet, for feature extraction in our models. This approach has the potential to improve model performance and efficiency.

F. Residual Attention U-Net

We propose implementing the Residual Attention U-Net to expand the comparison of U-Net-based architectures. This architecture has exhibited promising results in various computer

vision tasks and may contribute to advancing the state-of-the-art in kidney segmentation.

G. Investigation of the Dice Loss Function

Exploring the Dice loss function as an alternative to the Cross-Entropy and focal loss functions would provide further insights into its effectiveness in addressing the challenges specific to kidney segmentation.

By pursuing these future research directions, we anticipate significant advancements in the performance, robustness, and clinical applicability of the proposed models for kidney segmentation in MRI images using deep learning.

REFERENCES

- [1] N. M. Selby, P. J. Blankestijn, P. Boor, C. Combe, K.-U. Eckardt, E. Eikefjord, N. Garcia-Fernandez, X. Golay, I. Gordon, N. Grenier *et al.*, "Magnetic resonance imaging biomarkers for chronic kidney disease: a position paper from the european cooperation in science and technology action parenchima," *Nephrology Dialysis Transplantation*, vol. 33, no. suppl_2, pp. ii4-ii14, 2018.
- [2] K. Sharma, C. Rupprecht, A. Caroli, M. C. Aparicio, A. Remuzzi, M. Baust, and N. Navab, "Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease," *Sci. Rep.*, vol. 7, no. 1, p. 2049, May 2017.
- [3] Z. Zhao, H. Chen, and L. Wang, "A coarse-to-fine framework for the 2021 kidney and kidney tumor segmentation challenge," in *Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KiTS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*. Springer, 2022, pp. 53–58.
- [4] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *J Med Phys*, vol. 35, no. 1, pp. 3–14, Jan. 2010.
- [5] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang, X. Liu, J. Chen, H. Zhou, I. Ben Ayed, and H. Zheng, "Annotation-efficient deep learning for automatic medical image segmentation," *Nature Communications*, vol. 12, no. 1, p. 5915, Oct. 2021.
- [6] W. Brisbane, M. R. Bailey, and M. D. Sorensen, "An overview of kidney stone imaging techniques," *Nat Rev Urol*, vol. 13, no. 11, pp. 654–662, Aug. 2016.
- [7] J. L. Zhang, H. Rusinek, H. Chandarana, and V. S. Lee, "Functional mri of the kidneys," *Journal of magnetic resonance imaging*, vol. 37, no. 2, pp. 282–293, 2013.
- [8] D. C. Preston, "Magnetic Resonance Imaging (MRI) of the brain and spine: Basics," <https://case.edu/med/neurology/NR/MRI%20Basics.htm>, 2006, [URL revised 07/04/16], [Accessed 09-Jun-2022].
- [9] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022.
- [10] N. Siddique, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug 2018. [Online]. Available: <https://doi.org/10.1007/s13244-018-0639-9>
- [11] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82 031–82 057, 2021.
- [12] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *CoRR*, vol. abs/1704.06857, 2017. [Online]. Available: <http://arxiv.org/abs/1704.06857>
- [13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.
- [14] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S092523122100477X>
- [15] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018. [Online]. Available: <https://arxiv.org/abs/1804.03999>

- [16] A. Tiwari, "Chapter 2 - supervised learning: From theory to applications," in *Artificial Intelligence and Machine Learning for EDGE Computing*, R. Pandey, S. K. Khatri, N. Kumar Singh, and P. Verma, Eds. Academic Press, 2022, pp. 23–32. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780128240540000265>
- [17] H. Rezaatoghhi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [18] F. van Beers, A. Lindström, E. Okafor, and M. Wiering, "Deep neural networks with intersection over union loss for binary image segmentation," 02 2019.
- [19] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 237–242.
- [20] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 851–860, 2020.
- [21] E. Tiu, "Metrics to Evaluate your Semantic Segmentation Model," <https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2>, 2019, [Accessed 07-Feb-2022].
- [22] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. Blaschko, "Optimization for medical image segmentation: Theory and practice when evaluating with dice score or jaccard index," *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, 06 2020.
- [23] Y. Kurmi and V. Chaurasia, "Multifeature-based medical image segmentation," *IET Image Processing*, vol. 12, no. 8, pp. 1491–1498, 2018. [Online]. Available: <https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-ipr.2017.1020>
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [25] E. Koh, E. R. Walton, and P. Watson, "VIBE MRI: an alternative to CT in the imaging of sports-related osseous pathology?" *Br J Radiol*, vol. 91, no. 1088, p. 20170815, Mar. 2018.
- [26] N. M. Rofsky, V. S. Lee, G. Laub, M. A. Pollack, G. A. Krinsky, D. Thomasson, M. M. Ambrosino, and J. C. Weinreb, "Abdominal mr imaging with a volumetric interpolated breath-hold examination," *Radiology*, vol. 212, no. 3, pp. 876–884, 1999.
- [27] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (cycleGAN) to improve generalizability in ct segmentation tasks," *Scientific Reports*, vol. 9, no. 1, 2019, cited By :284. [Online]. Available: www.scopus.com
- [28] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, <http://www.deeplearningbook.org>.
- [29] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 424–432.
- [30] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," 2018.
- [31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.
- [32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [34] A. LUNTZ, "On estimation of characters obtained in statistical procedure of recognition," *Technicheskaya Kibernetika*, 1969. [Online]. Available: <https://cir.nii.ac.jp/crid/1573387449978580096>
- [35] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, pp. 1–20, 11 2019. [Online]. Available: <https://doi.org/10.1371/journal.pone.0224365>