# Impact of MR sequences choice on deep learning segmentation of muscles

Maylis Jouvencel, Hoai-Thu Nguyen, Magalie Viallon, P Croisille, Thomas Grenier

# Impact of MR sequences choice on deep learning segmentation of muscles

Maylis Jouvencel*, Hoai-Thu Nguyen*, Magalie Viallon†, Pierre Croisille†, Thomas Grenier‡
*Univ Lyon, UJM Saint-Etienne, INSA Lyon, UCB Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, F-42023, Saint-Etienne, France
†Department of Radiology, Centre Hospitalier Universitaire de Saint-Etienne, Université de Saint-Etienne, F-42055 Saint-Etienne, France
‡Univ Lyon, INSA Lyon, UCB Lyon 1, UJM Saint-Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Villeurbanne, France

***Abstract*—Medical image segmentation is a critical step for many medical studies. We address the problem of muscle segmentation on MRI images using Dixon sequences and explore the impact on the segmentation results when combining the four Dixon sequences available. Different combinations were put to test using two UNet-based architectures. One used an early fusion and input the images in the same encoder, while the other used late fusion, which learns the features from the images in separated encoders and then concatenates and decodes them as a whole. Our results show that the T1 water-only image is the most appropriate image for muscle segmentation in our database and that both early and late fusion approaches did not yield significantly different results. Thus, appropriate check of most adequate contrast to consider is feasible and recommended to exquisitely match to the observed population and the early fusion architecture appears to be the most efficient design to do so when dealing with such muscle segmentation task.**

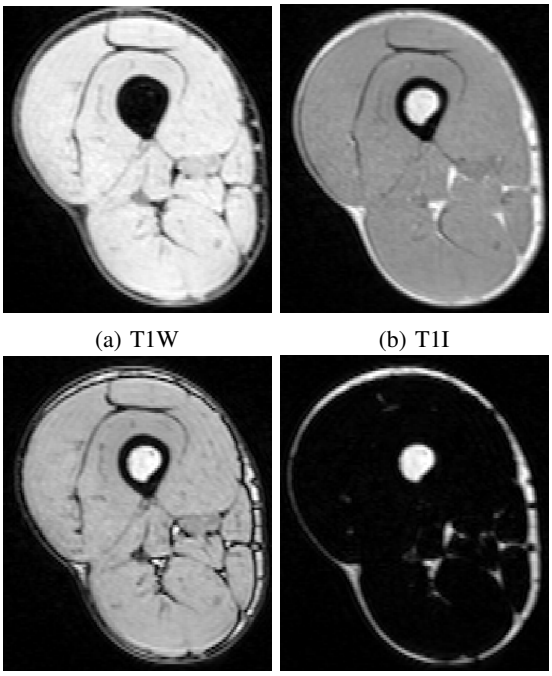
***Index Terms*—medical image segmentation, convolutional neural network, MRI**

## I. Introduction

Precise measurements of muscle volume are interesting for longitudinal studies regarding metabolism [1], the effects of physical effort, or diet on the body [2], [3]. Those volumes can be obtained with the segmentation of MRI images of muscles. As the manual 3D segmentation task is very tedious even for experts, automatic segmentation approaches are essential for longitudinal studies of muscles on 3D images on large cohorts of patients [4]. Deep learning methods have already proven themselves to be efficient tools for automatic segmentation [5], [6] and specifically for muscle segmentation [7].

However, the lack of annotated references due to the time and effort-consuming nature of the manual segmentation process is an obstacle to the development of those tools. Data augmentation [8] or other schemes [9], [10] can be used to tackle the issue. Nonetheless, we aim to investigate if using different modalities of the same MRI image, which have the same segmentation, can improve the results of the methods already implemented such as U-Net [5].

Our MRI images come from a Dixon acquisition sequence. Four types of T1-weighted images can be obtained using Dixon method, as described in [11]. Those images are Water only (later referred to as T1W), Fat only (T1F), In-phase (T1I), and Out-of-phase (T1O). Figure 1 shows an example of the Dixon images for our dataset. Previous studies [7], [12] used T1W images to segment muscles. [4] also showed that when using a registration method on the four images separately, T1W gives the best results. As the most efficient recent approaches for medical image segmentation are based on artificial neural networks [6], our work inspected whether it was still the case with a UNet-based network [5] and if information combined from the four types of images could also bring some improvements. The interest of such a study can be to save time, reduce the complexity of analysis and the size of the data handled by performing the segmentation task without certain image sequences.



(a) T1W (b) T1I

(c) T1O (d) T1F

Fig. 1: Example of the 4 sequences of a central slice from our dataset.

To do so, we first focused on a classical way of taking into account more than one input in networks by working with one channel per sequence. This is also known as *early fusion*

scheme. Then, we studied whether a more dedicated architecture, such as *late fusion* scheme, could be more appropriate to handle our MRI sequences. Several works have already studied the difference between those two types of network for different applications. Authors of [13] first used the technique for brain segmentation with T1-weighted, T2-weighted and fractional anisotropy (FA) images. The works [14], [15] also used it for brain tumor segmentation, with different types of MRI images. In [16], authors applied this method on Dixon type images to segment inter-vertebral discs, but directly used the four Dixon images altogether while we also try different combinations of those images, as detailed in II.

These works found that late fusion gave better results than early fusion. We hypothesize that this is not always true. Thus, our work intend to investigate statistically whether it is still the case for segmentation of athletes' muscles using different combinations of Dixon type images.

This paper is organized as follows: we first present the methods tested with the early and late fusion networks. We then describe our experiments by detailing our data-set and the metrics used to assess the segmentations obtained. In the last section, we present and discuss our results.

## II. Method

The input of the networks was a combination of, at most, the four Dixon image types. The combinations used were:

- each image type individually,
- all the pairs of image types (for example: T1W and T1I),
- the four images (T1W,T1I,T1O,T1F) altogether, later referred to as ALL,

In total, 11 combinations were experimented.

Two networks based on UNet [5] were employed. For the individual images, a regular five-stages UNet was used, similarly to [7]. When there were more than one type of input, two options were possible with either an early or a late fusion network, meaning the network has one or multiple encoders, respectively.

### A. Early fusion network

With early fusion UNet, two or more images are concatenated to form a single input of a single-encoder network, like in [17] and [18]. Figure 2 illustrates the network when combining T1W and T1I to form the input.

### B. Late fusion network

With late fusion UNet, each image is fed to a separate encoder, as inspired by [13], [15]. Therefore, the network has 2 or 4 encoders with the same architecture depending if 2 or 4 images are used as input. The features resulting from these encoders are fused at a later step, at the bottleneck of the network. Figure 3 illustrates this late fusion network when T1W and T1I are used as input. We notice that the number of parameters to be trained in the late fusion network is 1.5 and 2.5 times higher than for the early fusion network with 2 and 4 inputs, respectively.

## III. Experiments

### A. Dataset

We used a dataset of MR volumes collected from 48 athletes during the Tor des Géants Mountain-Ultra-Marathon (MUM) 2014, acquired for the study of [19]. For this dataset, the goal is to provide segmentation of the quadriceps. We have ground truth annotations which are manual segmentations provided by medical experts. The dataset can be separated into two groups, according to the segmented volumes available:

- 41 subjects who have annotations on both legs for the 2D central slice only,
- 7 subjects who have the whole right leg annotated (3D), and the left leg with only the central slice annotated.

Due to the small number of fully annotated 3D volumes, which production by medical experts is particularly demanding, and in order to have results statistically meaningful, we decided to work in 2D with central slices.

The 7 subjects of the latter group were used only for training, while the 41 other subjects were split: 26 subjects used for training, 10 used for validation, and 5 used for testing. The right and left legs of a subject were always in the same set. In terms of number of slices, it made a total of 96 slices (because there are 48 subjects) with 66 slices used for training, 20 for validation, and 10 for testing. In order to increase the number of training images, we used more images from the right leg of the 7 fully annotated subjects. If we note the index of the central slice used as $i$, all the slices at index $i+5$, $i+10$, $i-5$, and $i-10$ are also used for training. This enabled us to increase the training set to 94 slices, with $7 \times 4 = 28$ slices added. The training/testing process was repeated 40 times in a cross-validation scheme (for each time, both train and test sets are randomized) to obtain the results presented in part IV.

Before the extraction of the slices, the MR volumes were pre-processed. First, an N4 bias field correction was applied [20]. Then, the intensities were standardized on the first subject of the dataset. This pre-processing was done on the four sets of Dixon-type images. Figure 1 shows the four Dixon images after pre-processing for one subject.

### B. Implementation details

Our work is implemented using TensorFlow 2.6.0. To obtain the best results, each architecture's hyper-parameters were manually optimized using an exhaustive search on filter number (32 or 64), batch size (16 or 32), and batch norm (with or without). The best results were produced with a batch size of 16 samples, an Adam optimizer, and a learning rate of $l_r = 10^{-4}$. Other important networks parameters used are given in figures 2 and 3. The loss function used during training is the categorical cross-entropy. Finally, no post-processing was applied to the segmentations produced.

Training and inference were performed on an NVidia RTX A5000 GPU. On this GPU, training a network for one combination took around 7 minutes with early fusion, regardless of the number of modalities we concatenated. Training for late fusion took around 10 minutes with 2 modalities and 25

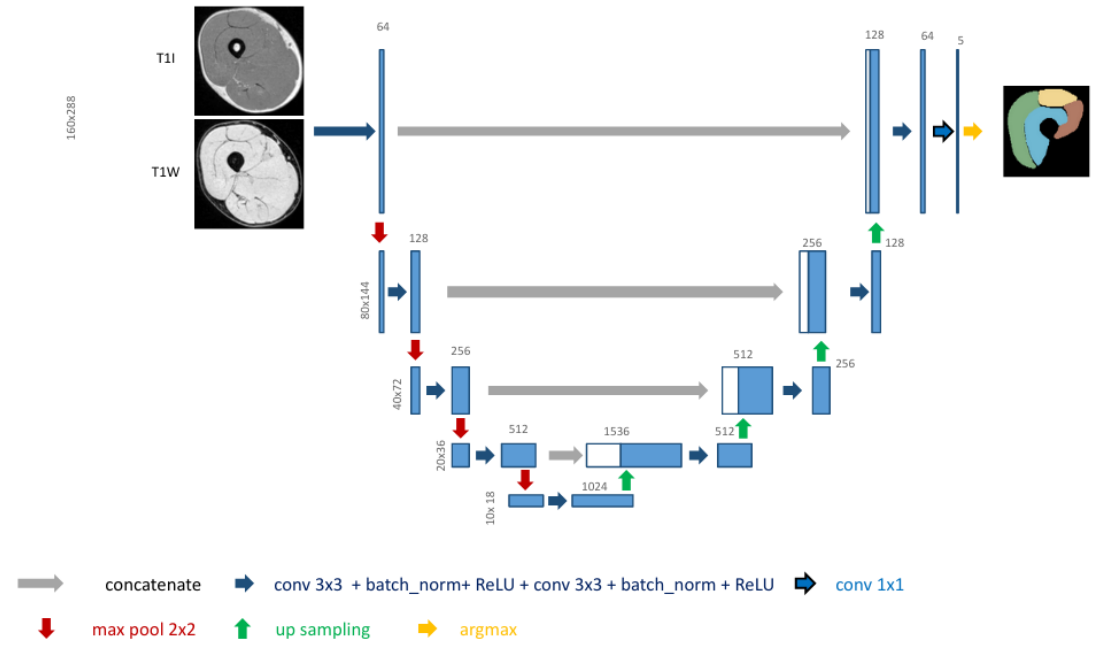



Fig. 2: Architecture of the early fusion UNet with T1W and T1I used as two channels for input.

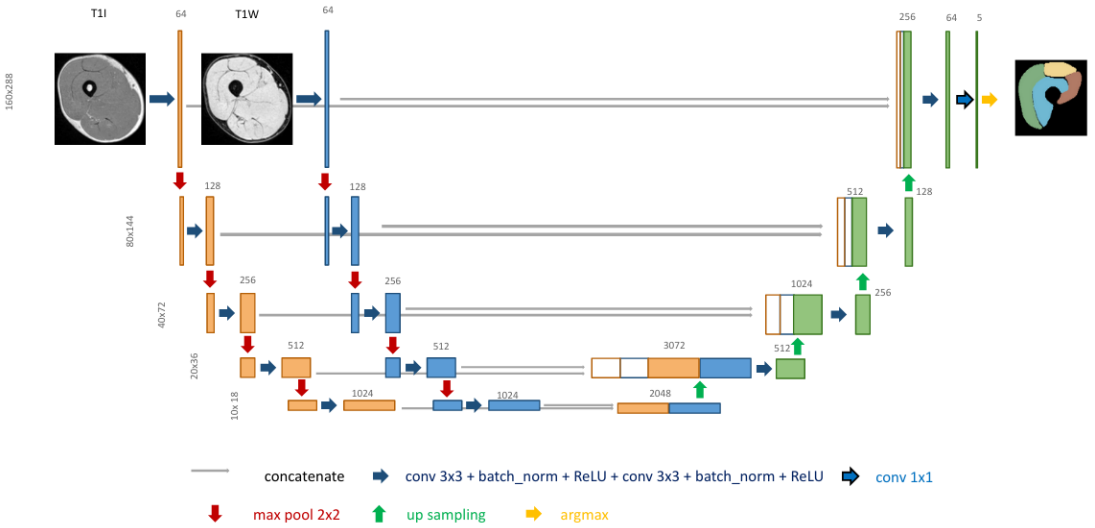



Fig. 3: Architecture of the late fusion UNet with T1W and T1I used as inputs for the two encoders.

minutes with 4 modalities. The test set inference usually took approximately 5 seconds for both architectures.

### C. Evaluation Metrics

Several evaluation metrics exist in order to evaluate the quality of a segmentation [21], among which we decide to use: Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Mean Absolute Distance (MAD).

DSC measures the overlap of a label mask predicted P with the label reference mask R, and is defined as :

$$DSC(R,P) = 2 \times \frac{|R \cap P|}{|R| + |P|} \quad (1)$$

The closer to 1 the DSC value is, the better the prediction. The segmentations we produce are multi-class with four labels corresponding to the four muscle heads that make up the quadriceps. The global DSC in this case is the averaged DSC of the four classes.

HD measures the largest distance between the surface of the prediction P and the reference R:

$$HD(R,P) = \max(d(R,P), d(P,R)) \quad (2)$$

$$\text{where } d(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

The closer to 0 the HD value is, the better the prediction. The global HD in a multi-class segmentation is largest HD value

among all classes.

Finally, MAD measures the mean distance between the surfaces of the reference and the predicted regions:

$$MAD(R,P) = \frac{d(R,P) + d(P,R)}{2} \quad (3)$$

$$\text{where } d(A,B) = mean_{a \in A} \min_{b \in B} \|a - b\|$$

The closer to 0 the MAD value is, the better the prediction.

The study of the results was mostly done by studying the DSC results. MAD and HD were taken into account for confirmation and for better understanding of the way errors are distributed (i.e. for further ad-hoc post-processing developments).

## IV. Results and Discussion

TABLE I: Quantitative evaluation of networks

| Combination | Method | DSC | HD (mm) | MAD (mm) |
|---|---|---|---|---|
| T1W | | $\mathbf{0.925 \pm 0.039}$ | $16.00 \pm 15.51$ | $1.89 \pm 1.88$ |
| T1I | | $0.900 \pm 0.069$ | $15.71 \pm 13.89$ | $2.18 \pm 1.52$ |
| T1O | | $0.892 \pm 0.081$ | $16.82 \pm 12.25$ | $2.58 \pm 2.47$ |
| T1F | | $0.848 \pm 0.089$ | $19.64 \pm 13.15$ | $3.48 \pm 2.73$ |
| T1W,T1I | early | $0.921 \pm 0.047$ | $14.84 \pm 14.00$ | $1.78 \pm 1.33$ |
| | late | $0.924 \pm 0.040$ | $\mathbf{14.16 \pm 12.73}$ | $\mathbf{1.71 \pm 1.41}$ |
| T1W,T1O | early | $0.921 \pm 0.041$ | $15.10 \pm 12.07$ | $1.81 \pm 1.46$ |
| | late | $0.920 \pm 0.045$ | $17.11 \pm 16.77$ | $1.97 \pm 1.80$ |
| T1W,T1F | early | $0.920 \pm 0.044$ | $15.04 \pm 12.43$ | $1.86 \pm 1.43$ |
| | late | $0.918 \pm 0.052$ | $18.13 \pm 17.61$ | $2.05 \pm 2.11$ |
| T1I,T1O | early | $0.903 \pm 0.073$ | $16.39 \pm 13.38$ | $2.20 \pm 1.74$ |
| | late | $0.904 \pm 0.072$ | $15.62 \pm 11.38$ | $2.05 \pm 1.46$ |
| T1I,T1F | early | $0.886 \pm 0.096$ | $16.01 \pm 11.41$ | $2.56 \pm 3.28$ |
| | late | $0.889 \pm 0.095$ | $18.98 \pm 17.70$ | $2.86 \pm 6.23$ |
| T1O,T1F | early | $0.884 \pm 0.105$ | $17.16 \pm 14.07$ | $2.99 \pm 5.94$ |
| | late | $0.886 \pm 0.102$ | $16.31 \pm 12.64$ | $2.70 \pm 3.10$ |
| ALL | early | $0.914 \pm 0.054$ | $15.51 \pm 12.12$ | $1.96 \pm 1.42$ |
| | late | $0.919 \pm 0.043$ | $15.13 \pm 11.43$ | $1.81 \pm 1.30$ |

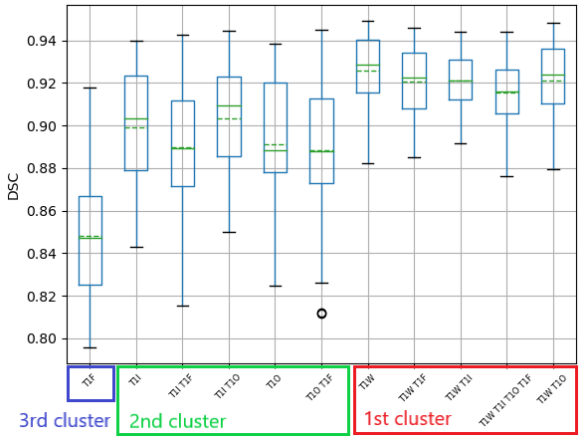
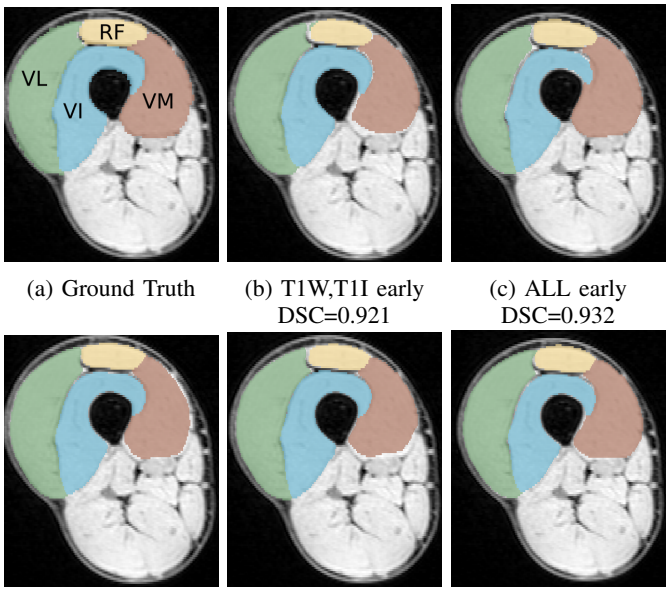



Fig. 4: Boxplot of DSC metric for early fusion UNet

Figure 5 illustrates the segmentation results of the late and early architectures on a test image slice using different input sequences. The visual comparison is quite satisfying and highlights similar mistakes near the separation between




(a) Ground Truth (b) T1W,T1I early DSC=0.921 (c) ALL early DSC=0.932

(d) T1W DSC=0.928 (e) T1W,T1I late DSC=0.929 (f) ALL late DSC=0.926

Fig. 5: Example of ground truth and predicted segmentations for the combinations : T1W, (T1W, T1I) and ALL with early and late UNet architectures.

vastus lateralis (VL) and rectus femoris (RF). Errors can also be found for vastus medialis (VM) in the area close to the femur.

More quantitatively, table I displays the mean metrics for both methods tested. The first observation we can make is that the DSC obtained are globally high: ten combinations out of the eleven have $DSC > 0.85$. We can also note that some combinations share similar metrics results, and we can in particular separate the set of combinations into three groups of combinations. T1W, (T1W, T1I), (T1W, T1O), (T1W, T1F) and ALL (i.e. (T1W, T1I, T1O, T1F) ) make up the first cluster with the best results. Indeed all have a high DSC with $DSC > 0.91$ and a low MAD, with $MAD < 2.0$ mm. T1I, T1O, T1F, (T1I,T1O), (T1I,T1F), (T1O,T1F) make up a second cluster with intermediate results: $DSC$ is in $[0.88, 0.91]$ and $MAD$ in $[2.0, 3.0]$. T1F alone makes up the last group with the lowest results: $DSC < 0.84$, and $MAD > 3.0$.

Figure 4 shows a visual confirmation of these observations. It indeed illustrates that for the DSC metrics, three distinct groups can be observed. The first cluster in particular has higher means and smaller standard deviations, which also signifies that the results are more robust.

In order to simplify the analysis, we then decide to focus on the results from the first cluster consisting of the five combinations previously mentioned, since they give the best results. We apply a two-sided Wilcoxon signed-rank test on the DSC values to see if there is a significant difference between the combinations tested. Tables II and III display the resulting P-values. Bold values have P<0.05 meaning that the difference is statistically significant.

We also want to see if the difference between early and late fusion is significant for each combination. We therefore perform another Wilcoxon test: for the applicable combinations, we study the difference between early and late fusion network. The results are displayed in table IV. Bold values have P$<$0.05.

The first result to note is that the best mean DSC and MAD are obtained for combinations which include T1W: those combinations make up the first cluster.

Table I shows that using T1W only gives the best mean DSC. However, the difference between this result and other combinations is significant only for some of them. Indeed P-value is greater than 0.05 when comparing T1W and (T1W, T1I) for both early and late fusion, which means that the difference between those two combinations is not statistically significant. It is therefore difficult to conclude on which one of T1W or (T1W, T1I) gives the best results, especially since the latter has the lowest MAD.

We then compare the results using the same combination but with either early or late fusion. From table I, one can note that metrics are close from early and late architectures, with no particular emerging rule. Only one combination, the one using the four images altogether, shows a significant difference between the two networks tested, which means a P-value$<$ $0.05$ (see table IV). Indeed, the late fusion network gives better result for the four images for the three metrics used. For the rest of the combinations, early and late fusion give similar results: Wilcoxon test shows that the difference between the two is not meaningful, with a P-value$>$ $0.05$. Consequently, it is not possible to conclude whether late fusion is better than early fusion for our dataset.

On another note, it must be reminded that the results were achieved for a dataset formed from a very specific type of subjects who are thin and muscular. While the poor results from T1F alone show that the fat information is not sufficient to produce segmentation, we can speculate that the results could be influenced positively on a dataset where the subjects are less athletic. Indeed, having more fat outside the muscles heads could have a positive impact on the results with combination including T1F, while having fat inside the muscles heads could on the contrary worsen the results.

TABLE II: Wilcoxon test for early fusion UNet on the first cluster of combinations

| | T1W,T1I | T1W,T1O | T1W,T1F | ALL |
|---|---|---|---|---|
| T1W | 0.0599 | **0.0497** | **0.0174** | **0.0009** |
| T1W,T1I | | 0.9893 | 0.9037 | **0.0187** |
| T1W,T1O | | | 0.7368 | **0.0275** |
| T1W,T1F | | | | **0.0231** |

TABLE III: Wilcoxon test for late fusion UNet on the first cluster of combinations

| | T1W,T1I | T1W,T1O | T1W,T1F | ALL |
|---|---|---|---|---|
| T1W | 0.1357 | **0.0009** | **0.0438** | **0.0066** |
| T1W,T1I | | **0.0348** | 0.0656 | **0.002** |
| T1W,T1O | | | 0.9893 | 0.4200 |
| T1W,T1F | | | | 0.5188 |

TABLE IV: Wilcoxon results comparing early and late fusion UNet

| Combination | P-value |
|---|---|
| T1W,T1I | 0.1466 |
| T1W,T1O | 0.8297 |
| T1W,T1F | 0.9678 |
| T1I,T1O | 0.3468 |
| T1I,T1F | 0.3678 |
| T1O,T1F | 0.7470 |
| ALL | **0.0497** |

## V. Conclusion

When applying muscle segmentation with UNet based networks on T1 weighted Dixon images where Water, Fat, In and Out of phase are extracted, we showed that the choice of the network's inputs is an important step. In particular, T1 water-only images are crucial, whether they are combined or not with other sequences. Thus, for our application study involving athletes, this T1 water-only image can be considered at first for such muscle segmentation task based on UNet architecture. However, taking into account the specificity of each database when choosing the input images could have interesting results, because some subjects with more fat between muscles could benefit from using other Dixon images.

We also found that for our dataset, comparing the two major schemes for mixing input images (the early and late fusion) with UNet architecture, didn't give significant results, contrary to what was demonstrated in previous works. Our work shows that the choice of early or late fusion for UNet depends on the given problem and data and, that the increase in the number of parameters for the late fusion network should also be taken into account. Working on other annotated datasets with a similar muscle segmentation task but more diverse subjects in terms of morphology and muscle tissue quality, or adding more slices to our own dataset could add more insight to the current results.

Given the results obtained, we can now study explanatory methods, as one based on interpretability, to investigate the importance of each input in the final segmentation without the necessity of performing exhaustive combinations of tests.

## References


[1] G. Râdegran, E. Blomstrand, and B. Saltin, "Peak muscle perfusion and oxygen uptake in humans: importance of precise estimates of muscle mass," Journal of applied physiology, vol. 87, 1999.

[2] R. Ross, J. Rissanen, H. Pedwell, and J. C. P. Shragge, "Influence of diet and exercise on skeletal muscle and visceral adipose tissue in men," Journal of applied physiology, vol. 81, 1996.

[3] D. B. Starkey, M. L. Pollock, Y. Ishida, M. A. Welsch, W. F. Brechue, J. E. Graves, and M. S. Feigenbaum, "Effect of resistance training volume on strength and muscle thickness," Medicine & Science in Sports & Exercise, vol. 28, pp. 1311–1320, 1996.

[4] B. Gilles, C. De Bourguignon, P. Croisille, G. Millet, O. Beuf, and M. Viallon, "Automatic segmentation for volume quantification of quadriceps muscle head: a longitudinal study in athletes enrolled in extreme mountain ultra-marathon," in ISMRM: International Society for Magnetic Resonance in Medicine, 2016.

[5] 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.

[6] 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.

[7] H.-T. Nguyen, P. Croisille, M. Viallon, S. Leclerc, S. Grange, R. Grange, O. Bernard, and T. Grenier, "Robustly segmenting quadriceps muscles of ultra-endurance athletes with weakly supervised U-Net," in International Conference on Medical Imaging with Deep Learning, London, 2019.

[8] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, 2019.

[9] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Proceedings of the IEEE, vol. PP, pp. 1–34, 07 2020.

[10] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," 2019.

[11] J. Ma, "Dixon techniques for water and fat imaging," Journal of Magnetic Resonance Imaging, vol. 28, no. 3, pp. 543–558, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.21492

[12] H.-T. Nguyen, P. Croisille, M. Viallon, C. De Bourguignon, R. Grange, S. Grange, and T. Grenier, "Robust multi-atlas MRI segmentation with corrective learning for quantification of local quadriceps muscles inflammation changes during a longitudinal study in athletes," in ISMRM: International Society for Magnetic Resonance in Medicine, Paris, 2018.

[13] D. Nie, L. Wang, Y. Gao, and D. Shen, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1342–1345, 2016.

[14] M. Aygün, Y. H. Şahin, and G. Ünal, "Multi modal convolutional neural networks for brain tumor segmentation," 2018. [Online]. Available: https://arxiv.org/abs/1809.06191

[15] N. Debs, T.-H. Cho, D. Rousseau, Y. Berthezène, M. Buisson, O. Eker, L. Mechtouff, N. Nighoghossian, M. Ovize, and C. Frindel, "Impact of the reperfusion status for predicting the final stroke infarct using deep learning," NeuroImage: Clinical, vol. 29, p. 102548, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158220303855

[16] J. Dolz, C. Desrosiers, and I. B. Ayed, "Ivd-net: Intervertebral disc localization and segmentation in mri with a multi-modal unet," ArXiv, vol. abs/1811.08305, 2018.

[17] J. Ding, P. Cao, H.-C. Chang, Y. Gao, S. H. S. Chan, and V. Vardhanabhuti, "Deep learning-based thigh muscle segmentation for reproducible fat fraction quantification using fat–water decomposition mri," Insights into Imaging, vol. 11, 2020.

[18] R. Amer, J. Nassar, D. Bendahan, H. Greenspan, and N. Ben-Eliezer, "Automatic segmentation of muscle tissue and inter-muscular fat in thigh and calf mri images," in MICCAI, 2019.

[19] H.-T. Nguyen, T. Grenier, B. Leporq, C. Le Goff, B. Gilles, S. Grange, R. Grange, G. Millet, O. Beuf, P. Croisille, and M. Viallon, "Quantitative Magnetic Resonance Imaging Assessment of the Quadriceps Changes during an Extreme Mountain Ultramarathon," Medicine and Science in Sports and Exercise, vol. 53, no. 4, pp. 869–881, Apr. 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02998029

[20] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," IEEE Transactions on Medical Imaging, vol. 29, no. 6, pp. 1310–1320, 2010.

[21] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," BMC medical imaging, vol. 15, p. 29, 2015. [Online]. Available: https://europepmc.org/articles/PMC4533825