The Impact of Fatty Infiltration on MRI Segmentation of Lower Limb Muscles in Neuromuscular Diseases: A Comparative Study of Deep Learning Approaches

Marc-Adrien Hostin, MS,[1,2]* ● Augustin C. Ogier, PhD,[2,3] Constance P. Michel, PhD,[1]

Yann Le Fur, PhD,[1] Maxime Guye, MD, PhD,[4] Shahram Attarian, MD,[5] Etienne Fortanier, MD,[5]

Marc-Emmanuel Bellemare, PhD,[2] and David Bendahan, PhD[1] ●

Background: Deep learning methods have been shown to be useful for segmentation of lower limb muscle MRIs of healthy subjects but, have not been sufficiently evaluated on neuromuscular disease (NDM) patients.
Purpose: Evaluate the influence of fat infiltration on convolutional neural network (CNN) segmentation of MRIs from NMD patients.
Study Type: Retrospective study.
Subjects: Data were collected from a hospital database of 67 patients with NMDs and 14 controls (age: 53 ± 17 years, sex: 48 M, 33 F). Ten individual muscles were segmented from the thigh and six from the calf (20 slices, 200 cm section).
Field Strength/Sequence: A 1.5 T. Sequences: 2D T_1-weighted fast spin echo. Fat fraction (FF): three-point Dixon 3D GRE, magnetization transfer ratio (MTR): 3D MT-prepared GRE, T2: 2D multispin-echo sequence.
Assessment: U-Net 2D, U-Net 3D, TransUNet, and HRNet were trained to segment thigh and leg muscles (101/11 and 95/11 training/validation images, 10-fold cross-validation). Automatic and manual segmentations were compared based on geometric criteria (Dice coefficient [DSC], outlier rate, absence rate) and reliability of measured MRI quantities (FF, MTR, T2, volume).
Statistical Tests: Bland–Altman plots were chosen to describe agreement between manual vs. automatic estimated FF, MTR, T2 and volume. Comparisons were made between muscle populations with an FF greater than 20% (G20+) and lower than 20% (G20−).
Results: The CNNs achieved equivalent results, yet only HRNet recognized every muscle in the database, with a DSC of 0.91 ± 0.08, and measurement biases reaching $-0.32\% \pm 0.92\%$ for FF, 0.19 ± 0.77 for MTR, -0.55 ± 1.95 msec for T2, and -0.38 ± 3.67 cm^3 for volume. The performances of HRNet, between G20− and G20+ decreased significantly.
Data Conclusion: HRNet was the most appropriate network, as it did not omit any muscle. The accuracy obtained shows that CNNs could provide fully automated methods for studying NMDs. However, the accuracy of the methods may be degraded on the most infiltrated muscles (>20%).
Evidence Level: 4.
Technical Efficacy: Stage 1.

Neuromuscular diseases (NMDs) are a broad group of diseases that all affect nerves or muscles. Progressive muscle and nerve damage leading to loss of motor function has been reported as a hallmark of NMDs.[1] The challenge related to the follow-up of these changes is to have access to specific and sensitive biomarkers, which could be used to

assess the natural history of a disease or evaluate the efficiency of a therapeutic strategy.[2] These potential biomarkers could also be helpful for a better understanding of the disease mechanisms and supportive diagnosis.[3] Over the last decade, fatty infiltration has been recognized as a ubiquitous phenomenon in many neuropathies and dystrophies.[4] Fatty infiltration is a progressive replacement of muscle tissue by fat, which can be identified using quantitative MRI approaches quantifying fat fraction (FF).[5] FF has been shown to be more sensitive to disease progression than clinical or myometric measurements.[6] Although their interest as biomarkers has been less recognized than FF, other types of acquisition, such as T2 mapping and MTR, have been used to study NMDs.[7–9] While the clinical potential of these MRI biomarkers has been widely documented, utilization in clinics has not been reached so far. This limited use is related to the availability of sequences and the variability of scans but also to the fact that biomarker quantification requires a preliminary step of delineation (segmentation) of the regions of interest (i.e. individual muscles). To benefit from the 3D nature of MRI, segmentation should be performed on a volume, that is, on all the slices along the proximo-distal axis.[10–12] Segmentation should distinguish each individual muscle given that disease extent can vary among different muscles.[13–17] It has been largely recognized that manual segmentation is not an option in this specific context.[18,19] The corresponding task is time-consuming and suffers from operator dependency.[20]

In the literature, to date, a limited number of fully or semi-automatic methods have been reported to be effective in reducing segmentation time.[18,19,21–26] However, the high inhomogeneity of fatty infiltration among individual muscles regarding patients and diseases is a major concern for those methods.[18,19] As a matter of example, in Charcot-Marie-Tooth (CMT) patients, FF can vary from 0% to 81%.[11] A semi-automated method based on contour propagation of a few manually segmented slices provided interesting results, but manual entry can still be considered prohibitive for large-scale clinical use.[27,28]

Recently, a few studies have assessed the potential of deep learning methods based on convolution neural networks (CNNs).[19,21–26] These studies have shown promising segmentation results on healthy subjects, but few have investigated severe fat infiltration. Three studies reported results in severely infiltrated patients.[21,26,29] Although each study reported a difference in outcomes between moderate and severe infiltrations, the evolution of CNN performance as a function of FF was not investigated. In addition, the potential of segmentation methods was evaluated based on similarity between manual and automatic segmentations but not on the validity of quantitative MRI biomarker measurements. To date, only Chen et al and Ding et al evaluated the accuracy of quantitative MRI biomarkers using fully automatic segmentation. However, the evaluation was performed on a

homogeneous base of lightly infiltrated subjects, as evidenced by their FF measurements.[23,24] Overall, the robustness of CNN-based segmentation methods with respect to the extent of fat infiltration remains unknown.

Among the large variety of neural networks dedicated to segmentation, only recent CNNs, known to be efficient for medical image segmentation, have been chosen to address this question.[30–33] The obvious choice was U-Net 2D, the reference CNN for medical image segmentation.[30] U-Net 3D, a variant of U-Net for processing volumes, was selected to compare the 2D and 3D approaches.[31] Two other networks, transUNet and HRNet, were selected, as they address the recognized shortcomings of U-Net. TransUNet allows evaluating the relation between the different elements of the image thanks to the presence of the vision transformer module.[32] HRNet overcomes the loss of information due to the compression of the image in the U-Net encoder.[33]

The objective of this study was to evaluate the performance of these four networks, as a function of fatty infiltration. To enrich the training and evaluation of CNNs, a database consisting of patients from three NMDs and controls was collected.

Subjects and Methods
Standard Protocol Approvals, Registrations, and Patient Consents
The study was approved by the local research committee and was conducted in conformity with the Declaration of Helsinki (version October 2013) and the Medical Research Involving Human Subjects Act. Prior written informed consent was obtained from all subjects. Each patient provided an informed consent for a retrospective analysis of the MR images recorded as part of the research protocol they volunteered for.

Subjects
Data were collected from a hospital database collecting the work of three previous studies, on three different diseases, from the reference center for NMD and ALS at the university hospital of La Timone.[12,34] The cohort consisted of 67 patients with NMD and 14 controls (age: 53 ± 17 years, sex: 48 M, 33 F). The patients included 29 familial amyloid polyneuropathies (FAP), 18 CMT diseases, and 20 chronic inflammatory demyelinating polyneuropathies (CIDP). There was no recruitment or inclusion processes, since the data were only collected from previous studies, which had their own recruitment/inclusion process. A few patients have been scanned several times and the whole set of images have been integrated in the training database, to ensure proper training of the CNNs. The corresponding MRI dataset was composed of 218 MRI volumes (112 thighs and 106 legs).

Briefly, familial amyloid polyneuropathy (FAP) is a rare genetic disorder with autosomal-dominant inheritance due to a mutation in the transthyretin (TTR) gene, which causes a rapid progressive polyneuropathy.[35] All subjects had a confirmed mutation in the TTR gene, with 25 symptomatic patients and 14 presymptomatic carriers. CMT disease is the most common cause of

hereditary neuropathy.[36] All patients from our cohort were genetically confirmed as CMT1A patients with a classic mutation in the PMP22 gene. The third type of patient was composed of chronic inflammatory demyelinating polyneuropathy (CIDP), an acquired immune-mediated neuropathy characterized by a sensory-motor impairment. All CIDP patients fulfilled the definite clinical and electrophysiological European Federation of Neurological Societies (EFNS)/Peripheral Nerve Society (PNS) criteria for CIDP.[37] No patient had any history of other neuromuscular condition. The control group was composed of individuals with no medical history of neuropathic or muscular disease.

MRI Acquisitions

MRI scans were recorded at 1.5 T (MAGNETOM Avanto, Siemens Healthineers, Erlangen, Germany) at the thigh and leg levels using a spine coil on the bottom and two flexible coils on the top of the lower limb. 2D T1-weighted fast spin echo (T1w) images (repetition time [TR] = 549 msec; time to echo [TE] = 11 msec; flip angle [FA] = 120°; bandwidth = 195 Hz/pixel; in-plane matrix size/voxel size = 320 × 320/0.68 × 0.68 mm^2; 20 slices [slice thickness = 10.00 mm]; slice gap = 5.00 mm) were acquired and used as anatomical images for the muscle segmentation process. In addition, a three-point Dixon 3D gradient-echo (GRE) sequence (TR = 22 msec; eight echoes with TE from 2.38 msec to 19.06 msec in steps of 2.80 msec; bandwidth = 1220 Hz/pixel; FA = 5°; matrix size = 128 × 128 × 36; resolution = 1.72 × 1.72 × 5.00 mm^3) was acquired and used to generate fat fraction maps as previously described.[27] MTR maps were generated from a 3D MT-prepared GRE (TR = 36 msec; TE = 3.50 msec; FA = 10°; bandwidth/pixel = 200 Hz/pixel; matrix size = 128 × 128 × 36; resolution = 1.72 × 1.72 × 5.00 mm^3; MT offset frequency = 1200 Hz; MT flip angle = 750°; handshaped MT pulse duration = 10.0 msec). T$_2$ maps were generated from the postprocessing of a 2D multispin-echo sequence (TR = 2500 msec; 16 echoes with TE from 8.7 msec to 139.2 msec in steps of 8.7 msec; FA = 180°; bandwidth = 454 Hz/pixel; in-plane matrix size/voxel size = 128 × 128/1.72 × 1.72 mm^2; 10 slices [slice thickness = 20.00 mm]; slice gap = 10.00 mm).

Ground Truth Segmentations

A total of 10 (respectively 6) individual muscles were delineated in the thigh (respectively in the leg). The ROIs for the thighs consisted of the following individual muscles: adductor (Ad), *biceps femoris* (BF), *gracilis* (Gr), *rectus femoris* (RF), *sartorius* (Sa), *semimembranosus* (SM), *semitendinosus* (ST), *vastus intermedius* (VI), *vastus lateralis* (VL), and *vastus medialis* (VM). For the leg, the following muscle groups were delineated: anterior compartment (AC), deep posterior compartment (DPC), *gastrocnemius lateralis* (GL), *gastrocnemius medialis* (GM), lateral compartment (LC), and *soleus* (So).

The segmentations were performed by neurologists who had participated in the various research protocols from which the data were gathered, and each had at least 5 years of segmentation experience (E.F., C.P.M.). The segmentations were also reviewed by another nonclinical operator with 3 years of experience in manual segmentation (M.-A.H).

Manual segmentation was performed on T$_1$w images. The raters only segmented a limited number of slices, depending on the

patient, and a semi-automatic method using a combination of diffeomorphic registrations was used to propagate these segmentations to the remaining slices.[27,28] The final segmentations were checked by the same observers, to correct the propagated masks if needed.

Implementation

CNN architecture appears in Fig. 1. U-Net 3D is not presented since its architecture is very similar to U-Net 2D, with 3D convolutions instead of 2Ds. The network is four-layered and the number of channels is 24/48/96/192 for the encoder, and the decoder is built symmetrically. The volumes were padded with replicated slices on the extremities to create a 320 × 320 × 24 volume, making it easier to divide the number of slices per two at each layer. The characteristics of the other neural networks are described in Fig. 1.

A Python 3.8.3 environment was used to implement the CNN training with PyTorch 1.11.0 Experiments were run on a Linux Xeon Silver Workstation (4214cpu@2.2 GHz–96 Gb) with a Nvidia GeForce RTX 3090 GPU.

The T$_1$w images were used for the CNNs training and consisted of 112/106 thigh/calf volumes for the 3D set and 2240/2120 thigh/calf images for the 2D set. The validation set represented 10% of the training set. Networks were individually trained using 10-fold cross validation. The loss function was the Dice loss, a standard function for image segmentation library for deep learning. The optimization algorithm was the PyTorch version of Adam.[38] Each network was trained with an early stopping strategy with patience of 10 epochs. No data augmentation was performed, as the addition of random rotation and shift did not result in improved performance of the CNNs.

Evaluation

The performance of each network was compared to the ground truth (manual segmentation) based on geometric similarity metrics and MRI biomarkers agreement. The geometric metric included the commonly accepted, that is, DSC, an index of segmentation quality ranging from 0 (no overlap) to 1 (total overlap). In addition, for each metric, the outlier rate (OR) and the rate of unidentified muscles (AR) were computed. The OR was calculated as the rate of metric values that are lower to the threshold value Q1-1.5*IQR with IQR = Q3-Q1, Q1 being the lower quartile and Q3 the upper quartile of the corresponding metric. The AR represents the ratio between the number of muscles volumes identified by operators and the number of muscles detected by the CNNs.

From the quantitative MRI maps (FF, MTR, T2), the values of each quantity were calculated as the average of the values of the corresponding map over the entire volume. For each metric, a prediction error score ΔX was computed to represent the difference between the measured values with the masks from automatic segmentation X$_a$ and ground truth X$_m$, where X stands for FF, MTR, T2 or volume.

Statistical Analysis

A benchmark analysis of the performance of the segmentation methods was conducted. This was based on two criteria: 1) average similarity between automatic and manual segmentations through DSC, OR and AR; and 2) average measurement bias of the MRI quantities (ΔFF, ΔMTR, ΔT2, and ΔV).
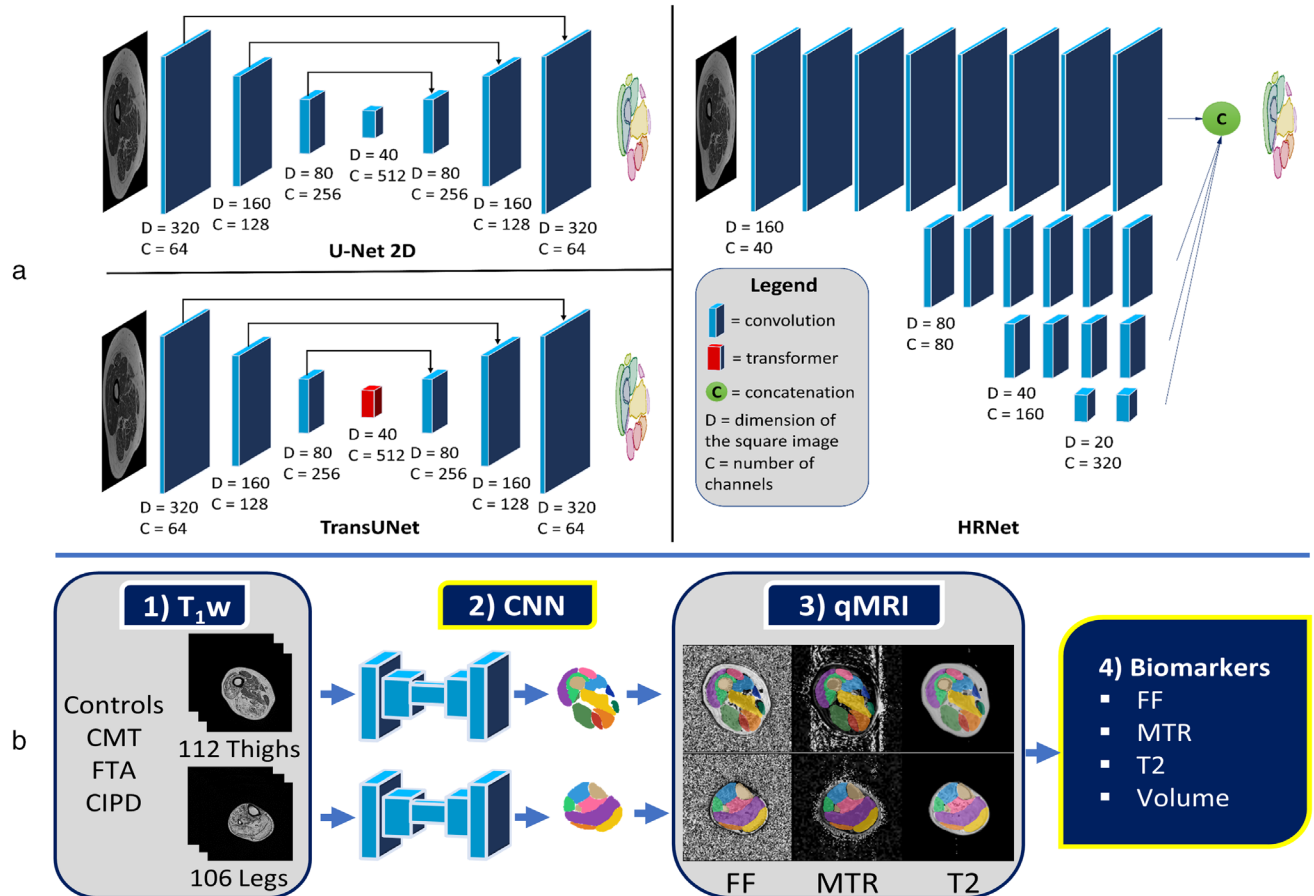
FIGURE 1: (a) Networks architecture including number of channels and number of convolutional layers with U-Net 2D, TransUNet, and HRNet. D: dimension of the square image, C: number of channels. (b) Processing pipeline composed of 1) training CNNs on T1w image database; 2) predicting segmentations with one of the four trained CNNs; 3) applying predicted segmentations to each qMRI maps to; 4) extract scores for each biomarker to compare them with those from manual segmentations.

Based on this comparison, the measurement bias of the best model was investigated regarding the degree of fatty infiltration, represented by the FF, using a Bland–Altman plot.

The impact of fat infiltration on performance was also investigated by studying the differences in accuracy between two subgroups of muscles with FFs less (G20−) and greater (G20+) than 20%. The 20% threshold was chosen since it was approximately the maximum value on which CNNs segmentation has been evaluated in former studies.[23,24] To test the difference in accuracy between G20+ and G20−, the distribution of samples was initially evaluated using the Shapiro–Wilk test. Differences were then assessed using non-parametric Wilcoxon pairwise tests or parametric Student's t-tests. The significance level was set at $P < 0.05$.

Results

CNN Training

Training time ranged from 2 hours (U-Net 2D) to 3 hours (HRNet) per network. To perform 10-fold cross validation, the total time was thus between 20 hours for U-Net 2D and 30 hours for HRNet.

Fat Infiltration Distribution

As illustrated in Fig. 2, the distribution of the biomarkers of interest was heterogeneous. Heterogeneity was described by the values of mean ± standard deviation, range, and coefficient of variation for each metric: FF (8.76 ± 7.68, [1.87, 64.34], 0.88) (%); MTR (48.01 ± 6.21, [9.69, 55.78], 0.13); T2 (55.47 ± 15.86, [34.72, 151.72]; 0.29) (msec).

Ground Truth vs. Predicted Segmentation

DSC, ΔFF (%), ΔMTR, ΔT2 (msec), and ΔV (cm³) values for each model and each muscle are presented in Fig. 3. For the whole dataset, the average values were 0.92 ± 0.05 (DSC), −0.28 ± 1.00% (ΔFF), 0.17 ± 0.83% (ΔMTR), −0.49 ± 2.11 msec (ΔT2), and −0.35 ± 16.83 cm³ (ΔV).

As can be seen in Fig. 3, DSC values were high, with the largest value obtained for *Ad* (0.95 ± 0.01) and the lowest for *GM* (0.86 ± 0.09). The smallest FF error was observed for *So* (0.32 ± 0.33) and the largest for *GM* (1.19 ± 1.16). ΔMTR reached the minimal value for *So* (0.28 ± 0.27) and the maximal for *Gr* (1.09 ± 0.98). Regarding the ΔT2 (msec), *So* was the muscle with the
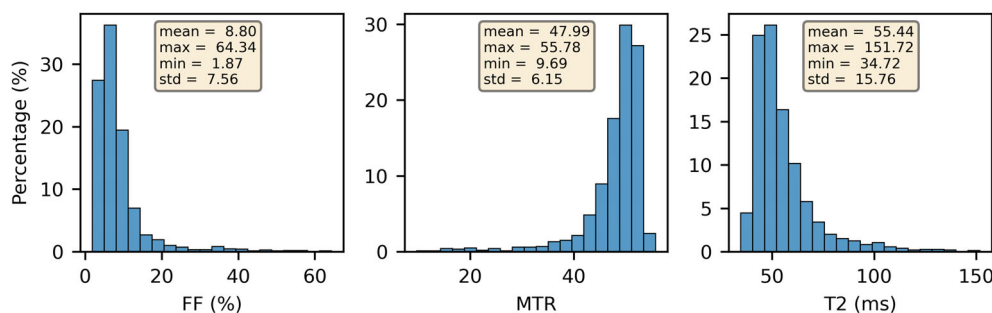
FIGURE 2: Distribution of FF (%), MTR, and T2 (msec) over our dataset composed of 96 thighs and 90 legs, that is, 1500 muscles.

smallest error (-0.81 ± 0.90), whereas the largest error was observed for *Gr* (-2.95 ± 2.39). The smallest volume error $\Delta V(cm^3)$ was found for *Gr* (3.37 ± 3.14) and the largest for *So* (15.66 ± 14.11).

The Pearson correlation coefficient between DSC and each biomarker metrics did not exceed -0.45.

Comparison of Network Performances
The performance of each network on the thigh and leg is presented for each metric in Table 1.

For the thigh, results ranged from [0.92 ± 0.05 (U-Net 3D), 0.93 ± 0.04 (HRNet)] for DSC, [$-0.32 \pm 0.82\%$ (U-Net 2D), $-0.37 \pm 0.91\%$ (HRNet)] for FF, [0.17 ± 0.86 (3D U-Net), 0.25 ± 0.76 (HRNet)] for MTR, [-0.62 ± 1.84 msec (HRNet), -0.55 ± 1.77 msec (2D U-Net)] for T2, and [-0.25 ± 3.18 cm^3 (HRNet), -0.11 ± 3.11 cm^3 (2D U-Net)] for volume.

For the leg, results ranged from [0.88 ± 0.08 (3D U-Net), 0.89 ± 0.11 (HRNet)] for DSC, [$-0.23\% \pm 0.93\%$ (HRNet), $-0.10\% \pm 1.12\%$ (transUNet)] for FF, [0.07 ± 0.68 (2D U-Net), 0.09 ± 0.77 (HRNet)] for
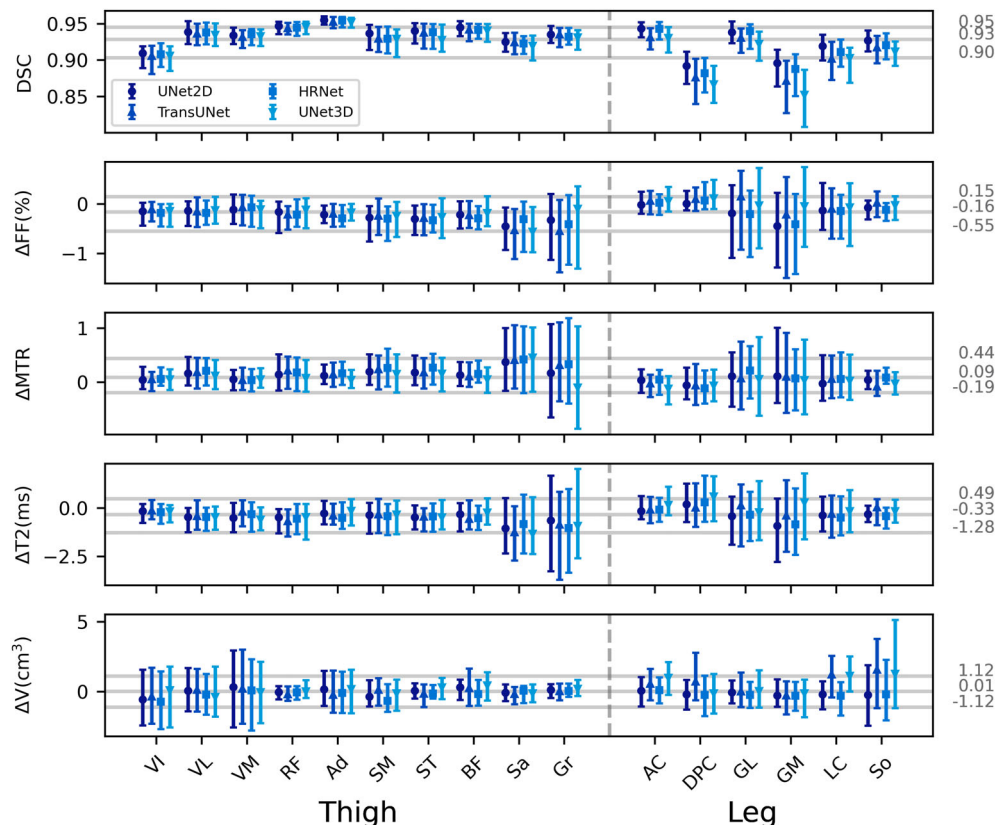

FIGURE 3: DSC, ΔFF (%), ΔMTR, ΔT2 (msec) and ΔV (cm^3) for each model and each muscle. Values are represented as the median with the first and third quartiles, respectively. The horizontal lines indicate the median values averaged for the whole set of CNNs and muscles (median), the averaged third quartile (upper) and the averaged first quartile (lower) (corresponding values are displayed on the right side). Thigh: Ad, *adductor*; BF, *biceps femoris*; Gr, *Gracilis*; RF, *rectus femoris*; Sa, *sartorius*; SM, *semimembranosus*; ST, *semitendinosus*; VI, *vastus intermedius*; VL, *vastus lateralis*; VM, *vastus medialis*. Calf: AC, anterior compartment; DPC, deep posterior compartment; GL, *gastrocnemius lateralis*; GM, *gastrocnemius medialis*; LC, lateral compartment; So, *soleus*.

TABLE 1. Average Evaluation Scores Over Thigh and Leg Muscles for HRNet, TransUNet, U-Net 2D, and U-Net 3D

Section	Network	DSC	ΔFF (%)	ΔMTR	ΔT2 (msec)	ΔV (cm³)	AR (%)	OR (%)
Thigh	HRNet	0.93 ± 0.04	-0.37 ± 0.91	0.25 ± 0.76	-0.62 ± 1.84	-0.25 ± 3.18	0.00	5.77
	TransUNet	0.92 ± 0.04	-0.34 ± 0.94	0.21 ± 0.78	-0.60 ± 1.99	-0.17 ± 3.56	2.28	6.68
	UNet2D	0.93 ± 0.04	-0.32 ± 0.82	0.20 ± 0.75	-0.55 ± 1.77	-0.11 ± 3.11	1.70	4.53
	UNet3D	0.92 ± 0.05	-0.34 ± 1.20	0.17 ± 0.86	-0.55 ± 2.12	-0.11 ± 3.47	4.04	4.76
Leg	HRNet	0.89 ± 0.11	-0.23 ± 0.93	0.09 ± 0.77	-0.42 ± 2.12	-0.63 ± 4.40	0.00	7.52
	TransUNet	0.88 ± 0.08	-0.16 ± 1.13	0.07 ± 0.94	-0.33 ± 2.73	0.44 ± 3.44	0.00	6.64
	UNet2D	0.89 ± 0.12	-0.17 ± 0.88	0.07 ± 0.68	-0.41 ± 2.16	-0.51 ± 3.45	3.89	7.48
	UNet3D	0.88 ± 0.08	-0.10 ± 1.12	0.07 ± 0.99	-0.04 ± 2.27	0.50 ± 3.45	1.39	6.39

MTR, $[-0.42 \pm 2.12$ msec (HRNet), -0.04 ± 2.27 msec (3D U-Net)] for T2, and $[-0.63 \pm 4.40$ cm³ (HRNet), 0.44 ± 3.44 cm³ (transUNet)] for volume.

Regarding identification error, HRNet reached an AR score of 0%, whereas this score ranged from 1.39% to 4.04% for the other networks (Table 1). The OR scores slightly varied between networks and were between 4.53% and 5.77% for the thigh and 6.39% and 7.52% for the leg. The rate of outlier was about 2% higher for the leg muscles as compared to the thigh muscles. Figure 4 shows an illustration of missing muscle segmentation, representing the AR, and examples of errors in muscle contour detection, representing the OR. The AR score is illustrated by an example of missing muscle segmentations in Fig. 4b. Similarly, the OR score is highlighted by examples of poor muscle contour detection in Fig. 4c,d.

Robustness to Fat Infiltration

The sensitivity of measurement errors to fat infiltration is illustrated by the Bland–Altman plot shown in Fig. 5. Considering that the whole set of networks performed similarly well and for the sake of clarity, Bland–Altman plots are only presented for HRNet, the CNN for which the whole set of muscles was identified.

As indicated by the confidence intervals, the reliability of the measurements ranged between $[-2.12, 1.49]$ (%) for ΔFF, $[-1.32, 1.7]$ for ΔMTR, and $[-4.36, 3.26]$ (msec) for ΔT2. The reliability for volume error was smaller with a larger confidence interval equal to $[-7.58, 6.81]$ (cm³). The mean error and standard deviation for each metric were consistently larger for the G20+ group than the G20 group (Table 2). In addition, the statistical study revealed that the increase in error was significant for each metric (P values $<1.00 \times 10^{-3}$).

Discussion

In this study, the performance of CNNs for automatic segmentation of individual muscles was evaluated. 2D U-Net, 3D U-Net, transUNet, and HRNet were selected from the state of the art to perform this task. The models were tested on a large database heterogeneous in degree of fat infiltration, composed of patients from three different NMDs and a population of controls. A comparison of the results was made according to the similarity of the predicted segmentations with the manual references (DSC, AR, and OR) as well as the agreement between the MRI quantities measured with the segmentations of the CNNs and the references (FF, MTR, T2, volume). The results of the best model, HRNet, were studied against the degree of infiltration, that is, FF. Finally, the muscle population was divided into two infiltration subgroups (G20−, G20+) to perform a statistical comparison of HRNet performance on each group.

The accuracy of the automatically predicted segmentations, illustrated by DSC values, was slightly better than the values reported in the literature using CNNs (from 0.77 to 0.93)[21,23,24,26] and similar to the values reported using non learning semi-automatic methods (0.90 ± 0.03).[27] It is noteworthy that for the latter method, manual segmentation of a few slices was needed as a preliminary step, and this prerequisite might be seen as a limitation for large clinical applications.[18] Our method was consistent, as the results obtained on our heterogeneous database in fatty infiltration (ranging from 1.87% to 64.34%) were comparable with the scores reported in the literature on a slightly infiltrated set of subjects (<20%).[24,32] Moreover, our results were superior to those obtained on a set of severely infiltrated subjects in the studies of Rohm et al (0.85 ± 0.08), Agosti et al (0.87), and Gadermayr et al. (0.88).[21,26,29] Although of interest, DSC cannot be considered as a standalone index for assessing the performance of a segmentation algorithm in a clinical context.
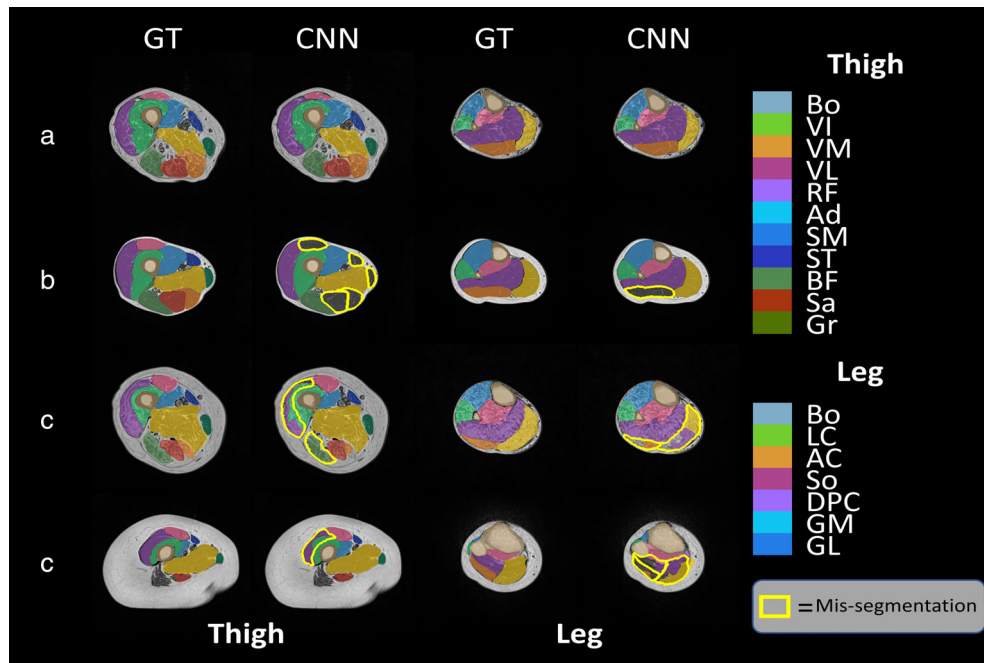
FIGURE 4: Examples of 2D U-Net-based segmentation for thigh and leg muscles showed in the axial plane. (a) Appropriate segmentation, (b) example of unidentified muscles. (c, d) Errors in contours detection. Thigh: Ad, *adductor*; BF, *biceps femoris*; Gr, *Gracilis*; RF, *rectus femoris*; Sa, *sartorius*; SM, *semimembranosus*; ST, *semitendinosus*; VI, *vastus intermedius*; VL, *vastus lateralis*; VM, *vastus medialis*. Calf: AC, anterior compartment; DPC, deep posterior compartment; GL, *gastrocnemius lateralis*; GM, *gastrocnemius medialis*; LC, lateral compartment; So, *soleus*.
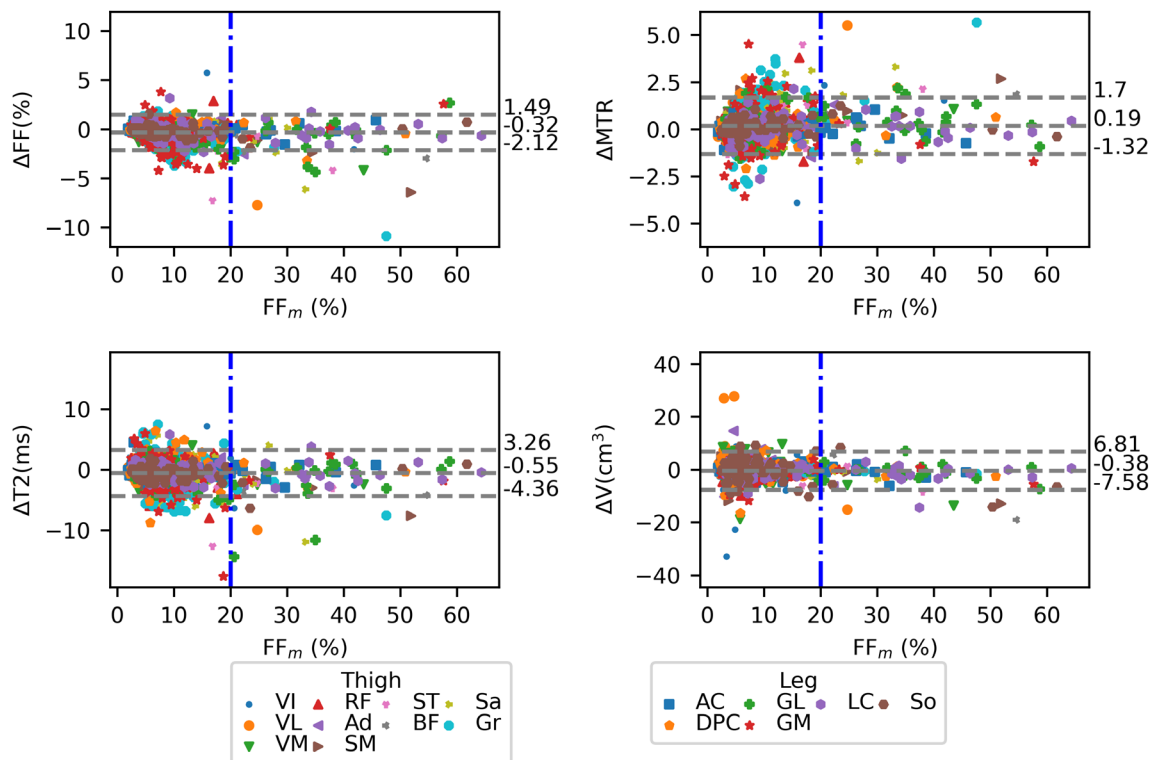

FIGURE 5: Bland–Altman type plots showing the biomarker errors obtained with HRNet segmentations with respect to fat fraction. Horizontal lines indicate the ±95% confidence interval of each measurement error. The vertical blue lines indicate the 20% delimitation between moderate and severe fat infiltration. Each symbol represents the error on one muscle. Thigh: Ad, *adductor*; BF, *biceps femoris*; Gr, *Gracilis*; RF, *rectus femoris*; Sa, *sartorius*; SM, *semimembranosus*; ST, *semitendinosus*; VI, *vastus intermedius*; VL, *vastus lateralis*; VM, *vastus medialis*. Calf: AC, anterior compartment; DPC, deep posterior compartment; GL, *gastrocnemius lateralis*; GM, *gastrocnemius medialis*; LC, lateral compartment; So, *soleus*.

TABLE 2. Average Errors Obtained With HRNet Segmentations on Muscles With FF < 20% (G20) and FF > 20% (G+)

Metric	G20−	G20+	*P* value
DSC	0.92 ± 0.04	0.87 ± 0.09	4.48E-15
ΔFF (%)	−0.27 ± 0.77	−1.09 ± 2.11	2.43E-04
ΔMTR	0.17 ± 0.73	0.56 ± 1.25	1.00E-03
ΔT2 (msec)	−0.48 ± 1.82	−1.69 ± 3.20	4.98E-04
ΔV (cm^3)	−0.19 ± 2.91	−1.90 ± 4.63	3.89E-04

While DSC informs about the overlapping between a predicted and a manually segmented mask, one must quantify clinically useful indices such as FF, MTR, and T2. Rather counter-intuitively, DSC values in our evaluation were poorly correlated with biomarkers quantification errors. In other words, a high DSC value would not be indicative of a reduced error regarding other metrics of interest. For example, a segmentation mask with a high DSC could properly cover a muscle region. However, if it also covers some other voxels around the muscle, that is, intermuscular fat, then the FF could be highly biased. Similarly, if the DSC is low, but the segmentation is in the central part of a muscle with a homogeneous infiltration, quantification of biomarkers would not be biased, and the corresponding results would be close to those from the ground truth. Thus, from a clinical outcome perspective, the most relevant performance indicator in segmentation studies may be biomarker quantification.

In a clinical context and more specifically in the field of neuromuscular disorders, CNNs are expected to provide a high-quality segmentation of individual muscles, which could be used to compute MRI biomarkers with a very high accuracy. This is of utmost importance if one intends to use MRI biomarkers to follow-up the natural history of muscle diseases or to assess the efficiency of a therapeutic strategy in a short time window. The corresponding errors quantified in the present study were relatively low, −0.3% ± 1.0% for FF, 0.2 ± 0.8 for MTR, and −0.55 ± 1.94 msec for T2. Of interest, the errors were lower than the changes reported over a 12-month period in CMT1A (FF: 1.1% ± 2.4%, T2: 1.4 ± 2.6 msec, MTR: 1.1 ± 2.4).[7] This result clearly indicates that CNN-based segmentation could be used to compute MRI biomarkers of interest and characterize subtle changes in longitudinal follow-up studies. This would imply that the infiltrated images would be correctly segmented as well.

The accuracy of quantitative MRI biomarker quantification based on fully automatic segmentation methods has been poorly evaluated in the literature. Only two studies have reported FF quantification in individual muscles using U-Net-based segmentation and the corresponding assessment was performed in a limited number of patients, that is, 24 and 4, respectively.[23,24] While Ding et al reported a 0.17% systematic bias, Chen et al reported a confidence interval of [0.56%, 0.49%] for the thigh, and [0.71%, 0.84%] for the leg muscles.[23,24] FF errors computed in our entire database were slightly larger than theirs with a systematic bias of 0.28% and a CI of [2.12%, 1.49%]. However, it should be kept in mind that the infiltration range in our database (<64%) was much larger than those in the quoted studies (<20%) and this is likely to have a detrimental effect on the quality of segmentations. By selecting only muscles with 0%–20% infiltration, the biomarker estimation results (systematic bias: 0.20%, CI: [1.38%, 0.976%]) were closer to the values reported by Ding et al and Chen et al.[24,32]

Comparative analysis between systematic errors obtained in muscles with FF values below (G20−) and above 20% (G20+) conducted in our database indicated that the segmentation and accuracy of the corresponding biomarker were negatively affected by high FF values. This effect was particularly noticeable on volume, where the error was increased 10 times on severely infiltrated muscles. This distorting effect of FF could be related to the fact that fat infiltration could affect the visibility of muscle boundaries.

The volumes of muscles segmented by CNN were lower on average than the corresponding volumes calculated from manual segmentations. In other words, the volume of CNN segmentations was significantly underestimated on the most infiltrated muscles.

Detailed inspection of the individual MR images in our cohort suggests that the FF value is not the only factor responsible for the quantification bias. The pattern of fatty infiltration also appears to play a role. Although this aspect warrants further study, sparse infiltration would not prevent the detection of muscle contours and would not bias the quantification of muscle volume. On the contrary, a more compact infiltration would have a more detrimental effect.

In terms of geometric efficiency, the different CNNs performed almost equally well, with minimal differences

regarding DSC values. All the CNNs, apart from the HRNet, failed to identify some muscles. The non-identification corresponds to the fact that the CNN associates the absent muscles to the background of the image, thus an improvement track could be explored in this direction. Since the identification of individual muscles is essential for the correct quantification of biomarkers, this study identifies HRNet as the most appropriate network for the segmentation of muscle images.

Limitations

This study was performed with data from a single center and a single scanner. It would be relevant to compare the performance of the CNNs on another database, from another center, and on other neuromuscular diseases (eg facioscapulohumeral muscular dystrophy). As the accuracy of deep learning methods is highly dependent on the nature of the training data, a transfer learning approach might be required to achieve the same results.[39]

Many neural networks could have been used in this study. Among the great variety of CNNs, we have chosen U-Net, the standard for medical image segmentation, as well as variants using a transformer module and 3D processing. To diversify our approach, we have chosen to test a different architecture from the encoder–decoder scheme with HRNet. We believe that the choice of these four networks was sufficient to evaluate the effect of fat infiltration on the automatic segmentation. Testing other architectures may be done in the future, but would be beyond the scope of this study.

The reliability of our approach was assessed from a comparative analysis with manual segmentation. Although the manual segmentation strategy was carefully detailed in guidelines, previous comparative analyses have indicated that DSC values computed for segmentations performed by different observers ranged from 0.80 to 0.95.[27] We considered that this variability regarding manual segmentations was of interest given that it provided an additional source of heterogeneity which might be learned by the CNNs. However, this necessarily implies that without absolute ground truth, the goal of a fully automated segmentation method should be to achieve an accuracy that matches the inter-operator variability.

Conclusion

All four networks tested in this study provided high-quality segmentations on FAP, CMT, and CIDP patients. That indicates/illustrates they could be used for accurate quantification of biomarkers. Although we identified a biasing effect of fat infiltration on biomarker accuracy, it was still acceptable compared with the 12-month patient biomarker trends, demonstrating the potential of follow-up studies.

Data Availability Statement

The principal author (PA) of the present manuscript has full access to the data used for the present manuscript and takes full responsibility of the data, the analyses and interpretation, and the conduct of the research. The PA also certifies that he has the right to publish all data, separate and apart from the guidance of any sponsor. The code is being formatted to be made public. It can be requested from the PA.

References

1. Morrison BM. Neuromuscular diseases. Semin Neurol 2016;36: 409-418.

2. Fischer D, Bonati U, Wattjes MP. Recent developments in muscle imaging of neuromuscular disorders. Curr Opin Neurol 2016;29(5):614-620.

3. Carlier PG, Reyngoudt H. The expanding role of MRI in neuromuscular disorders. Nat Rev Neurol 2020;16(6):301-302.

4. Mercuri E, Pichiecchio A, Allsop J, Messina S, Pane M, Muntoni F. Muscle MRI in inherited neuromuscular disorders: Past, present, and future. J Magn Reson Imaging off J Int Soc Magn Reson Med 2007;25(2): 433-440.

5. Bray TJ, Chouhan MD, Punwani S, Bainbridge A, Hall-Craggs MA. Fat fraction mapping using magnetic resonance imaging: Insight into pathophysiology. Br J Radiol 2017;91(1089):20170344.

6. Güttsches AK, Rehmann R, Schreiner A, et al. Quantitative muscle-MRI correlates with histopathology in skeletal muscle biopsies. J Neuromuscul Dis 2021;8(4):669-678.

7. Morrow JM, Sinclair CD, Fischmann A, et al. MRI biomarker assessment of neuromuscular disease progression: A prospective observational cohort study. Lancet Neurol 2016;15(1):65-77.

8. Sinclair C, Morrow J, Miranda M, et al. Skeletal muscle MRI magnetisation transfer ratio reflects clinical severity in peripheral neuropathies. J Neurol Neurosurg Psychiatry 2012;83(1):29-32.

9. Willcocks R, Arpan I, Forbes S, et al. Longitudinal measurements of MRI-T2 in boys with Duchenne muscular dystrophy: Effects of age and disease progression. Neuromuscul Disord 2014;24(5):393-401.

10. Janssen BH, Voet NB, Nabuurs CI, et al. Distinct disease phases in muscles of facioscapulohumeral dystrophy patients identified by MR detected fat infiltration. PLoS One 2014;9(1):e85416.

11. Gaeta M, Mileto A, Mazzeo A, et al. MRI findings, patterns of disease distribution, and muscle fat fraction calculation in five patients with Charcot-Marie-tooth type 2 F disease. Skeletal Radiol 2012;41: 515-524.

12. Bas J, Ogier AC, Le Troter A, et al. Fat fraction distribution in lower limb muscles of patients with CMT1A: A quantitative MRI study. Neurology 2020;94(14):e1480-e1487.

13. Ansari B, Salort-Campana E, Ogier A, et al. Quantitative muscle MRI study of patients with sporadic inclusion body myositis. Muscle Nerve 2020;61(4):496-503.

14. Heskamp L, van Nimwegen M, Ploegmakers MJ, et al. Lower extremity muscle pathology in myotonic dystrophy type 1 assessed by quantitative MRI. Neurology 2019;92(24):e2803-e2814.

15. Hooijmans M, Niks E, Burakiewicz J, et al. Non-uniform muscle fat replacement along the proximodistal axis in Duchenne muscular dystrophy. Neuromuscul Disord 2017;27(5):458-464.

16. Chung K, Suh B, Shy M, et al. Different clinical and magnetic resonance imaging features between Charcot–Marie–tooth disease type 1A and 2A. Neuromuscul Disord 2008;18(8):610-618.

17. Lareau-Trudel E, Le Troter A, Ghattas B, et al. Muscle quantitative MR imaging and clustering analysis in patients with facioscapulohumeral muscular dystrophy type 1. PLoS One 2015;10(7):e0132717.

18. Ogier AC, Hostin MA, Bellemare ME, Bendahan D. Overview of MR image segmentation strategies in neuromuscular disorders. Front Neurol 2021;12:255.

19. Gadermayr M, Disch C, Müller M, Merhof D, Gess B. A comprehensive study on automated muscle segmentation for assessing fat infiltration in neuromuscular diseases. Magn Reson Imaging 2018;48:20-26.

20. Barnouin Y, Butler-Browne G, Voit T, et al. Manual segmentation of individual muscles of the quadriceps femoris using MRI: A reappraisal. J Magn Reson Imaging 2014;40(1):239-247.

21. Rohm M, Markmann M, Forsting J, Rehmann R, Froeling M, Schlaffke L. 3D automated segmentation of lower leg muscles using machine learning on a heterogeneous dataset. Diagnostics 2021; 11(10):1747.

22. Guo Z, Zhang H, Chen Z, et al. Fully automated 3D segmentation of MR-imaged calf muscle compartments: Neighborhood relationship enhanced fully convolutional network. Comput Med Imaging Graph 2021;87:101835.

23. Chen Y, Moiseev D, Kong WY, Bezanovski A, Li J. Automation of quantifying axonal loss in patients with peripheral neuropathies through deep learning derived muscle fat fraction. J Magn Reson Imaging 2021;53:1539-1549.

24. Ding J, Cao P, Chang HC, Gao Y, Chan SHS, Vardhanabhuti V. Deep learning-based thigh muscle segmentation for reproducible fat fraction quantification using fat–water decomposition MRI. Insights Imaging 2020;11(1):1-11.

25. Cheng R, Crouzier M, Hug F, et al. Automatic quadriceps and patellae segmentation of MRI with cascaded U2-net and SASSNet deep learning model. Med Phys 2021;49(1):443-60.

26. Agosti A, Shaqiri E, Paoletti M, et al. Deep learning for automatic segmentation of thigh and leg muscles. Magn Reson Mater Phys Biol Med 2021;35:1-17.

27. Ogier AC, Heskamp L, Michel CP, et al. A novel segmentation framework dedicated to the follow-up of fat infiltration in individual muscles of patients with neuromuscular disorders. Magn Reson Med 2020;83(5): 1825-1836.

28. Ogier A, Sdika M, Foure A, Le Troter A, Bendahan D. Individual muscle segmentation in MR images: A 3D propagation through 2D non-linear registration approaches. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*: IEEE; 2017. p 317-320.

29. Gadermayr M, Li K, Müller M, et al. Domain-specific data augmentation for segmenting MR images of fatty infiltrated human thighs with neural networks. J Magn Reson Imaging 2019;49(6):1676-1683.

30. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference*, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18: Springer; 2015. p 234-241.

31. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference*, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19: Springer; 2016. p 424-432.

32. Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. ArXiv Preprint 2021;ArXiv: 2102.04306.

33. Sun K. High-resolution representations for labeling pixels and regions. ArXiv Preprint 2019;ArXiv190404514.

34. Fortanier E, Ogier AC, Delmont E, et al. Quantitative assessment of sciatic nerve changes in Charcot–Marie–tooth type 1A patients using magnetic resonance neurography. Eur J Neurol 2020;27(8):1382-1389.

35. Ando Y, Nakamura M, Araki S. Transthyretin-related familial amyloidotic polyneuropathy. Arch Neurol 2005;62(7):1057-1062.

36. Barreto LCLS, Oliveira FS, Nunes PS, et al. Epidemiologic study of Charcot-Marie-tooth disease: A systematic review. Neuroepidemiology 2016;46(3):157-165.

37. EFNS JTF of the, PNS T. European Federation of Neurological Societies/peripheral nerve society guideline on management of chronic inflammatory demyelinating polyradiculoneuropathy: Report of a joint task force of the European Federation of Neurological Societies and the peripheral nerve society–first revision. J Peripher Nerv Syst 2010; 15(1):1-9.

38. Kingma DP, Ba J. Adam: A method for stochastic optimization. ArXiv Preprint 2014;ArXiv14126980.

39. Ribani R, Marengoni M. A survey of transfer learning for convolutional neural networks. *2019 32nd SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T) 2019 Oct 28*: IEEE; 2019. p 47-57.