Automatic segmentation of ultrasound images of gastrocnemius medialis with different echogenicity levels using convolutional neural networks

Automatic segmentation of ultrasound images of gastrocnemius medialis with different echogenicity levels using convolutional neural networks

Francesco Marzola[1], Nens van Alfen[2], Massimo Salvi[1],Bruno De Santi[1],
Jonne Doorduin[2] and Kristen M. Meiburger[1]

***Abstract*— The purpose of this study was to develop an automatic method for the segmentation of muscle cross-sectional area on transverse B-mode ultrasound images of gastrocnemius medialis using a convolutional neural network(CNN). In the provided dataset images with both normal and increased echogenicity are present. The manually annotated dataset consisted of 591 images, from 200 subjects, 400 relative to subjects with normal echogenicity and 191 to subjects with augmented echogenicity. From the DICOM files, the image has been extracted and processed using the CNN, then the output has been post-processed to obtain a finer segmentation. Final results have been compared to the manual segmentations. Precision and Recall scores as mean $\pm$ standard deviation for training, validation, and test sets are 0.96 ± 0.05, 0.90 ± 0.18, 0.89 ± 0.15 and 0.97 ± 0.03, 0.89 ± 0.17, 0.90 ± 0.14 respectively. The CNN approach has also been compared to another automatic algorithm, showing better performances. The proposed automatic method provides an accurate estimation of muscle cross-sectional area in muscles with different echogenicity levels.**

I. INTRODUCTION

Ultrasound imaging (US) has been extensively used in the study of both normal and pathological skeletal muscles to extract quantitative information about muscle anatomy and morphology [1]. Magnetic resonance imaging(MRI) and computed tomography (CT) are also used for these tasks but US has the benefit of being inexpensive, safe, noninvasive and portable[2].

Among the muscle parameters that can be studied with US, there is cross-sectional area (CSA), muscle thickness, muscle echogenicity typically reported as the mean gray scale intensity, fascicle length, and pennation angle. These parameters are usually extracted manually by clinicians resulting in time-consuming and error-prone tasks[3]. These downsides give rise to the need for computer-aided approaches. Several automatic algorithms have been developed for the measurement of muscle thickness [4], CSA or other parameters both for transverse and longitudinal acquisition processes[5, 6].

Neuromuscular diseases (NMD) are often studied using US thanks to its ability to distinguish healthy and pathological muscles. Pathological muscles are subject to fatty replacement and increased presence of connective tissue, yielding an augmented echointensity that has been found to correlate with disease staging[7]. The correct staging of neuromuscular diseases through US requires an experienced sonographer but, even between experts, there is low accordance in the evaluation of the stage of an NMD[8]. Several systems have been proposed for the evaluation of the severity of neuromuscular diseases, but they all require manual interaction from the clinician for the selection of a region of interest, which is undesirable due to potential manual error that brings along and the time required for this operation[9]. The correct segmentation of CSA is a critical step in the evaluation of the muscle condition. Although algorithms for the automatic segmentation of muscles in B-mode ultrasound images have been proposed [10], as far as our knowledge goes, no one focuses on muscles with different echogenicity level making them not practical in evaluating subjects with pathological conditions [11].

We propose an automatic algorithm for the segmentation of B-mode ultrasound images of the gastrocnemius medialis with different echogenicity levels based on convolutional neural networks(CNN). This algorithm might be the starting point for a broader approach to automatise the whole process for the staging of an NMD from the segmentation of the muscle to its classification.

II. MATERIALS AND METHODS

A. Dataset description and imaging equipment & protocol

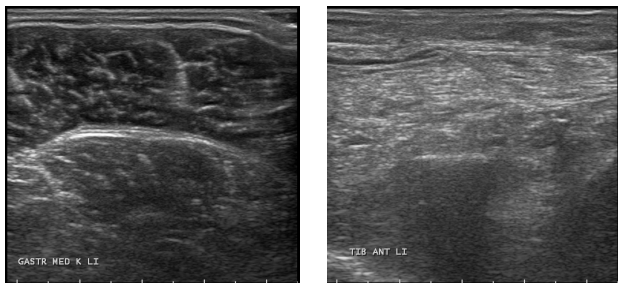
The dataset has been provided by the Department of Neurology, Radboud University Medical center. It has been acquired with the approval of the Research Ethics Committee (REC) of the Faculty of Science of the Radboud University. It comprises 591 images divided between 136 subjects with normal and 64 with augmented echogenicity. Each image has been acquired with ESAOTE® equipment, with gain set at 50% and depth at 44*mm* and anonymized using DicomCleaner®. For each subject, 2 or 3 images have been acquired. Subjects have been divided between *healthy* and *pathological* through echogenicity level. The mean gray value for each subject has been averaged among the 2-3 acquired images on a manually selected ROI. Subjects with normal echogenicity (*z-score* < 2) have been classified as healthy, while the ones with abnormal echogenicity (*z-score* > 2) as pathological. Example images can be seen in Figure 1. A summary of the subjects' characteristics can be seen in Table I, where *N*, *H*, *W* refer to *Number of subjects*, *Height* and *Weight*.

In particular, 400 images are related to healthy subjects while 191 images are related to pathological ones. Manual segmentations for the training and testing phases of the

[1]Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy;
[2]Department of Neurology, Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Center, Nijmegen, The Netherlands

TABLE I
SUBJECTS CHARACTERISTICS

	N	Sex(m/f)	Age(y)	H(cm)	W(kg)
Healthy	136	82/53	52 ± 21	175 ± 15	76 ± 20
Pathological	64	36/21	65 ± 21	168 ± 24	76 ± 25



(a) Healthy subject (b) Pathological subject

Fig. 1. Examples of images related to healthy and pathological subjects

algorithm have been provided by an operator with 2 years of experience in medical imaging with prior training from an expert with more than 10 years of experience in ultrasound image processing.

B. Algorithm summary

Our goal is to obtain the segmentation of the cross-sectional area of the gastrocnemius medialis starting from the original image without any interaction from a manual operator.

To achieve this result, the original DICOM image has been cropped using coordinates obtained from the DICOM header to obtain an image that the CNN could process. The resulting segmentation is then post-processed to obtain smoother boundaries. The final result is overlaid to the original image and various parameters can be calculated for the segmented area.

C. Network architecture

The architecture of our choice in this work has been the Unet with ResNet18 as encoder. The Unet architecture has proven to be a reliable solution in the segmentation of medical images[12]. This net comprises a down-sampling(encoder) section to extract features from the image and an up-sampling(decoder) section to reconstruct the original image. To obtain the features of the image the encoder of our choice has been the ResNet18 that has achieved good performances in classifying the ImageNet dataset[13]. This encoder exploits residual blocks, depicted in Figure 2, to obtain a very deep model while maintaining an acceptable model complexity to have reduced training time and a smoother loss surface[14].

D. Dataset splitting and data augmentation

Annotated images have been divided into training, validation and test sets. This division results in percentages of the total dataset of about 70% for the training set and 15% for both the validation and test sets.

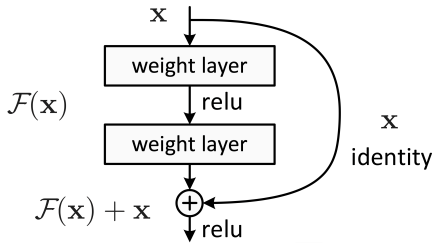


Fig. 2. The residual block splits in two the connections from its input. One connection goes through different convolutions and functions, while the other, named identity, connects directly to the output of these operations. The two outputs are then added together[13]

Data augmentation, is a critical step to achieve satisfactory performances and to ensure that the trained model can generalize to real-world data[15]. In this work we applied online data augmentation, the transformations applied to the original images are the following: rotation $\pm 5^{\circ}$, vertical shift (10% of image height), horizontal shift(10% of image width), horizontal flipping and shear.

E. Network implementation and training

The network we used for this project has been developed exploiting the Tensorflow platform using Keras API. The model architecture we used in this model is available in the Python library "Segmentation Models" [16]. For the weights initialization, we adopted the concept of transfer learning [17, 18]. The ResNet18 encoder weights of our Unet have been downloaded from (`https://github.com/fchollet/deep-learning-models`).

We have chosen the *Adam* optimizer as the optimization technique to minimize the loss function. We used default parameters implemented in Keras (learning rate=0.001, beta1=0.9, beta2=0.999, epsilon=$1e^{-08}$). As a loss function, we applied *binary cross-entropy* since only two classes are represented in our segmentation, the CSA and background. Performances of our model during training have been evaluated in terms of *intersection/union* in the training and validation sets.

Training has been performed with a mini-batch size of 32 images. Loss has been evaluated on the validation set for each epoch, patience was set at 8 epochs, and the model reached the early-stopping criterion at epoch 46. Total training time has been of 7 hours on an i5 based laptop with 16Gb of RAM.

F. Post-processing net output

The segmentation resulting directly from the prediction of the CNN is accurate. Despite that, in some cases, different connected areas are recognized as muscle or the muscle segmentation presents holes in it. To avoid these problems a light post-processing step has been implemented, this step can be visualized in Figure 3. If several connected areas are recognized as muscle, they are considered as possible candidates,

(a) Net output (b) Post-processed output

Fig. 3. Post-processing step visualization

the one with the largest area is selected as the right one. The outer boundary is smoothed using morphological operations (opening and closing with *disk* structural element of radius 10 pixels) to resemble physiological muscle structure in a more accurate way. These operations have also the purpose to fill small gaps that might be present in the CNN output.

III. RESULTS AND BENCHMARKING

A. Network performances

The proposed method has been evaluated in MATLAB® environment among the three sets derived from the initial dataset. Segmentation performance has been evaluated through the following scores[19]:

$$Precision = TP/TP + FN \quad (1)$$

$$Recall = TP/TP + FN \quad (2)$$

$$BFscore = 2 * precision * recall / precision + recall \quad (3)$$

In eq. 1 and 2 *precision* and *recall* are pixel based metrics. In eq. 3 *precision* and *recall* are referred to the contour of the segmented area. When evaluating contours, a true positive is found when the pixels on the automatic segmentation boundary and the pixels on the ground truth boundary have a distance under a certain threshold. In this work, this threshold has been set to $1mm$. The CNN correctly segmented 100% of the images. Results are summarized in table II in which all results are expressed as *mean* $\pm$ *standard deviation* and for each set the first row consider the entire set, while the second and the third focus only on healthy or pathological subjects. Two examples of final results can be seen in Figure 4.

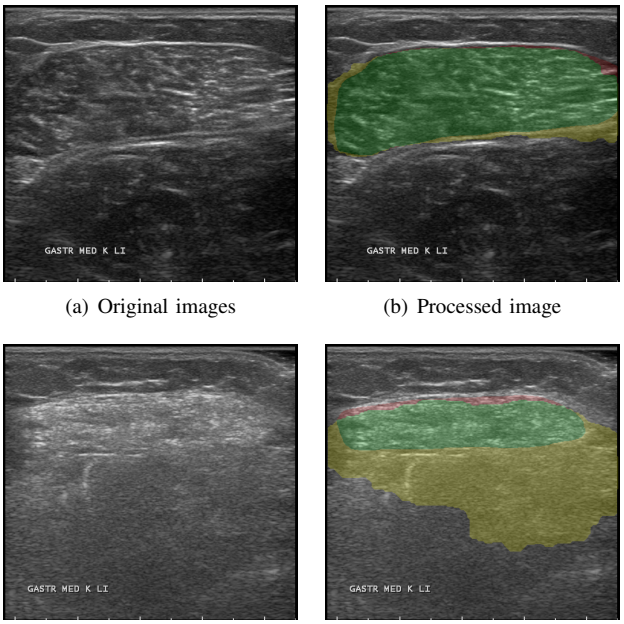
B. Comparison with a heuristic approach

The results produced by the CNN have been compared to the ones resulting from the application of another automatic segmentation approach, based on a heuristic approach developed on a different dataset(TRAnsverse Ultrasound Muscle Analysis - TRAMA)[6]. This algorithm is based on the recognition of the aponeuroses candidates with a multiscale filter and a heuristic search to detect the superior and inferior aponeuroses. We performed the comparison on the test set. We evaluated the *segmentation success rate* on all images and, the previously defined metrics, in images correctly segmented with both methods. The results are depicted in Table III.

TABLE II

EVALUATION METRICS FOR CNN ON DIFFERENT SUBSETS

	Precision	**Recall**	**BFscore**
Training	0.96 ± 0.05	0.98 ± 0.03	0.95 ± 0.06
Healthy	0.97 ± 0.03	0.98 ± 0.02	0.96 ± 0.06
Pathological	0.95 ± 0.06	0.97 ± 0.03	0.94 ± 0.06
Validation	0.90 ± 0.18	0.89 ± 0.17	0.78 ± 0.19
Healthy	0.89 ± 0.20	0.88 ± 0.19	0.78 ± 0.20
Pathological	0.90 ± 0.13	0.91 ± 0.10	0.81 ± 0.16
Test	0.89 ± 0.15	0.90 ± 0.14	0.74 ± 0.21
Healthy	0.93 ± 0.09	0.90 ± 0.12	0.78 ± 0.16
Pathological	0.80 ± 0.19	0.89 ± 0.18	0.65 ± 0.26



(a) Original images (b) Processed image

(c) Original images (d) Processed image

Fig. 4. Final processing result visualization. The first row shows a good result while the second a problematic case. The color scheme is the following, True Positive: Green, False Positive: Yellow, False Negative: Red, True Negative: not highlighted

IV. DISCUSSION AND CONCLUSIONS

In this study, we demonstrated the feasibility of using deep convolutional neural networks to segment transverse B-mode ultrasound images of gastrocnemius medialis. The performance of the proposed CNN is consistent along with the different subsets. The performance drop in the validation and test sets shows some overfitting, but the ability of the network to generalise on new images is not compromised. For both training and test sets, performances are higher for the images relative to healthy patients than for the one relative to pathological ones. This was highly expectable due to the lower contrast of the aponeuroses of pathological muscles. The whole algorithm takes under 2 seconds to give the segmentation of the muscle, which is compatible with clinical practice.

TABLE III
EVALUATION METRICS FOR TRAMA ALGORITHM

	S. Rate	Precision	Recall	BFscore
Test	0.78	0.73 ± 0.37	0.73 ± 0.34	0.56 ± 0.35
Healthy	0.87	0.87 ± 0.26	0.84 ± 0.27	0.70 ± 0.25
Pathological	0.62	0.35 ± 0.38	0.44 ± 0.44	0.18 ± 0.28

Comparing the proposed method to a heuristic approach the advantages of a deep learning based method are clear. TRAMA algorithm has defined constraints such as the need to define two aponeuroses that limit its ability to generalise when applied to a new dataset rising errors in 22% of images and having reduced performances, with respect to the CNN, in the images that it can process.

The proposed method could serve as the first step in the classification of neuromuscular diseases from ultrasound images. Limitations of this study are the unavailability of manual segmentations from an expert physician and a lack of cross-validated results from the CNN training that was not performed due to the very high computational cost.

Future work can be in the direction of addressing these limitations, improving the post-processing phase to refine segmentation and avoid errors in muscle detection. Another straightforward advancement would be the usage of this segmentation approach on a longer pipeline that would have as its final goal to provide a classification of the muscle to offer a second opinion to the physician.

REFERENCES

[1] Cristina Caresio et al. "Muscle echo intensity: reliability and conditioning factors". In: *Clinical Physiology and Functional Imaging* 35.5 (2015), pp. 393–403. DOI: 10.1111/cpf.12175.

[2] Peter Augat and Felix Eckstein. "Quantitative Imaging of Musculoskeletal Tissue". In: *Annual Review of Biomedical Engineering* 10.1 (2008), pp. 369–390. DOI: 10.1146/annurev.bioeng.10.061807.160533.

[3] Jean K. Mah and Nens van Alfen. "Neuromuscular Ultrasound: Clinical Applications and Diagnostic Values". In: *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques* 45.6 (2018), pp. 605–619. DOI: 10.1017/cjn.2018.314.

[4] P. Han, Y. Chen, and L. et al. Ao. "Automatic thickness estimation for skeletal muscle in ultrasonography: evaluation of two enhancement methods." In: *BioMed Eng OnLine* 12.6 (2013). DOI: 10.1186/1475-925X-12-6.

[5] Cristina Caresio et al. "Fully Automated Muscle Ultrasound Analysis (MUSA): Robust and Accurate Muscle Thickness Measurement". In: *Ultrasound in Medicine and Biology* 43.1 (2017), pp. 195–205. ISSN: 0301-5629. DOI: https://doi.org/10.1016/j.ultrasmedbio.2016.08.032.

[6] Massimo Salvi et al. "Transverse Muscle Ultrasound Analysis (TRAMA): Robust and Accurate Segmentation of Muscle Cross-Sectional Area". In: *Ultrasound in Medicine and Biology* 45.3 (2019), pp. 672–683. ISSN: 0301-5629. DOI: https://doi.org/10.1016/j.ultrasmedbio.2018.11.012.

[7] Nens van Alfen and Jean K. Mah. "Neuromuscular Ultrasound: A New Tool in Your Toolbox". In: *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques* 45.5 (2018), pp. 504–515. DOI: 10.1017/cjn.2018.269.

[8] Craig M. Zaidman et al. "Quantitative muscle ultrasound detects disease progression in Duchenne muscular dystrophy". In: *Annals of Neurology* 81.5 (2017), pp. 633–640. DOI: 10.1002/ana.24904.

[9] Philippe Burlina et al. "Automated diagnosis of myositis from muscle ultrasound: Exploring the use of machine learning and deep learning methods". In: *PLOS ONE* 12.8 (Aug. 2017), pp. 1–15. DOI: 10.1371/journal.pone.0184059.

[10] Xin Chen et al. "Automatic Tracking of Muscle Cross-Sectional Area Using Convolutional Neural Networks with Ultrasound". In: *Journal of Ultrasound in Medicine* 38.11 (2019), pp. 2901–2908. DOI: 10.1002/jum.14995.

[11] Zeynettin et al. Akkus. "A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow". In: *Journal of the American College of Radiology* 16.9, Part B (2019), pp. 1318–1328. ISSN: 1546-1440. DOI: https://doi.org/10.1016/j.jacr.2019.06.004.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[13] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[14] Hao Li et al. *Visualizing the Loss Landscape of Neural Nets*. 2017. arXiv: 1712.09913 [cs.LG].

[15] Z. et al Hussain. "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks". In: *AMIA Annu Symp Proc.* Apr 16 (2018), pp. 979–984.

[16] Pavel Yakubovskiy. *Segmentation Models*. https://github.com/qubvel/segmentation_models. 2019.

[17] N. Tajbakhsh et al. "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" In: *IEEE Transactions on Medical Imaging* 35.5 (May 2016), pp. 1299–1312. ISSN: 1558-254X. DOI: 10.1109/TMI.2016.2535302.

[18] Jason Yosinski et al. *How transferable are features in deep neural networks?* 2014. arXiv: 1411.1792 [cs.LG].

[19] G Csurka, D. Larlus, and F. Perronnin. *What is a good evaluation measure for semantic segmentation?* 2013.