Abdominal muscle segmentation from CT using a convolutional neural network

Edwards, Ka'Toria, Chhabra, Avneesh, Dormer, James, Jones, Phillip, Boutin, Robert, et al.

Ka'Toria Edwards, Avneesh Chhabra, James Dormer, Phillip Jones, Robert D. Boutin, Leon Lenchik, Baowei Fei, "Abdominal muscle segmentation from CT using a convolutional neural network," Proc. SPIE 11317, Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, 113170L (28 February 2020); doi: 10.1117/12.2549406

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Abdominal muscle segmentation from CT using a convolutional neural network

Ka'Toria Edwards [1], Avneesh Chhabra [2], James Dormer [1],
Phillip Jones [2], Robert D Boutin [3], Leon Lenchik [4], Baowei Fei [1,2*]

[1]*Department of Bioengineering, University of Texas at Dallas, Richardson, TX*
[2]*Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX*
[3]*Department of Radiology, UC Davis, CA*
[4]*Department of Radiology, Wake Forest School of Medicine, Winston-Salem, NC*
**E-mail: bfei@utdallas.edu, website: https://fei-lab.org*

ABSTRACT

CT is widely used for diagnosis and treatment of a variety of diseases, including characterization of muscle loss. In many cases, changes in muscle mass, particularly abdominal muscle, indicate how well a patient is responding to treatment. Therefore, physicians use CT to monitor changes in muscle mass throughout the patient's course of treatment. In order to measure the muscle, radiologists must segment and review each CT slice manually, which is a time-consuming task. In this work, we present a fully convolutional neural network (CNN) for the segmentation of abdominal muscle on CT. We achieved a mean Dice similarity coefficient of 0.92, a mean precision of 0.93, and a mean recall of 0.91 in an independent test set. The CNN-based segmentation method can provide an automatic tool for the segmentation of abdominal muscle. As a result, the time required to obtain information about changes in abdominal muscle using the CNN takes a fraction of the time associated with manual segmentation methods and thus can provide a useful tool in the clinical application.

Keywords: Muscle imaging, Image segmentation, Deep Learning, Muscle Segmentation, CT, Convolutional Neural Networks

INTRODUCTION

Computed tomography (CT) has been widely used in the clinics. CT is useful for routine patient evaluation, diagnosis of diseases, and monitoring changes in body composition. A recent study using abdominal CT scans reported age-related decreased skeletal muscle volume of 1.5 cm^2 per year and attenuation, or strength, reduction of 1.5 Hounsfield units (HU) per year[1]. Decreased skeletal muscle mass, referred to as sarcopenia, has been associated with increased morbidity and mortality related conditions, such as trauma, breast, colorectal, and lung malignancies. Sarcopenia can be fully diagnosed and monitored using CT imaging. Analysis of changes in muscle mass can be used to create optimal treatment plans.

Studies show that sarcopenia is a predictor of poor surgical outcomes for individuals that have experienced trauma. The predictive value for perioperative and long-term outcomes is what makes CT imaging so valuable for the treatment planning for patients suffering from sarcopenia[2, 3]. Further study of the progression of sarcopenia, through CT image analysis, can aid in the development of generalizable treatments for age-related and disease-related muscle loss. As a result, physicians may be able to improve active life expectancy in older people, and lead to substantial health-care savings and improved quality of life. Overall, the muscle loss associated with aging, sarcopenia, and chronic diseases is detectable and quantifiable using CT scans. Fully utilizing the ability to manipulate CT to track changes in muscle can greatly improve patient outcomes for those suffering from muscle loss related diseases[4].

There has been recent literature focused on body tissue segmentation around the abdomen and pelvis regions using deep learning and machine learning techniques. A CNN-based model has been created that enables accurate automated

Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, edited by Andrzej Krol, Barjor S. Gimi, Proc. of SPIE Vol. 11317, 113170L · © 2020 SPIE · CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2549406

segmentation of multiple tissues on pelvic CT images, with promising implications for body composition analysis. The deep CNN approach has the ability to achieve high accuracy when compared to manual segmentations, by a specialist, as the reference standard.[5-7] Such works are gaining scientific attention due to the fact that automatic segmentation of skeletal muscle on CT scans can be challenging because of the high variability between individuals. There are several factors that contribute to the variability of CT images. Shape of the individual, relative position within the CT machine, and unique characteristics among patients have the influence to introduce variability to the images which are being segmented. Until now, issues like these prevented automatic segmentation techniques from producing results comparable to that of trained specialists. In addition, similarities in texture, other neighboring muscles, uniqueness in the effects of particular muscular disorders, and varying intramuscular fat are sources of variability that limit the capabilities of current CNN models. Ensuring a diverse training data set is the only way to decrease the effects of these sources of variability between individuals and create a more robust CNN for CT image muscle segmentation.

A fully convolutional neural network segmentation method provides a viable alternative to manual segmentation approaches due to the consistent nature of the networks. In order to account for the vast array of potential CT images, many different patients with varying anatomy could be included in the data to train the network and create a robust model[8]. Small training data sets with little variability are the main limitation of the deep CNN automatic segmentation solution at this time. However, if the CNN is properly trained and validated, it can be very useful. For example, as a patient receives routine imaging during the course of treatment, abdominal images can be automatically segmented, and the muscle volume can be calculated and tracked automatically. Furthermore, muscle textural features could be assessed to evaluate the differences among various patients, which may serve as prognostic markers during the course of their disease management. Creating classification for textural features and changes in muscle mass are ways that CNNs can assist radiologists with characterization and analysis of disease. In this work, we investigated the ability of a fully convolutional neural network to segment skeletal abdominal muscle using CT slices. Then we tested its precision and accuracy with respect to manual segmentation performed by a specialist.

METHODS

Data Set Acquisition and Pre-processing

The CT imaging data was obtained on 33 adult patients (18-75 years). The slice thickness was 5.0 mm, with the slices ranging from the L2 level to lesser trochanters. An example can be seen in Figure 1A. A single CT slice from the L3 level of 39 additional patients were also included to increase the variability in the training data. The abdominal muscles were manually segmented for each CT image by a trained reader under the supervision of a fellowship-trained musculoskeletal radiologist. This resulted in a binary mask for each image for each patient that delineated the abdominal muscle of interest from the rest of the CT image and was considered the ground truth for the method (Figure 1B). An example of the original CT superimposed with the binary mask is show in Figure 1C. There are a total of two classes: background and muscle, where background was defined as any area that was not classified as skeletal abdominal muscle by the trained reader.

Figure 1. (A) CT slice and (B) manual segmentation for one patient. (C) The overlay of the segmentation on the CT slice.

In this work, we are using a supervised deep learning method where the machine learning task is to identify which part of the CT image is skeletal abdominal muscle based on example input-output pairs provided by the trained reader under the supervision of a fellowship trained radiologist. The CT images and segmentations were loaded into MATLAB (MathWorks, Inc., Natick, MA, USA). Each slice was 512 × 512 pixels in size. Hounsfield unit values from the CT images above 127 and below -128 were set to -128 to reduce the amount of information fed into the CNN, as has been shown to improve CNN segmentation for soft tissue segmentation in CT images[9]. Then, the images were rescaled to unsigned 8-bit portable network graphics files to reduce computational load on the neural network.

Patients and their corresponding images were separated into three groups: training, validation, and independent testing data. This ensured that patients which were used to train the CNN were not used again in the validation data designed to validate the CNN model or for final testing. The distribution of patients is shown in Table I. In the CNN, we did not make any additional parameters to account for patients known to be experiencing muscle loss. This ensured a more robust network that will contain independent predictive characteristics based solely on the segmentations provided by the specialist.

Table I. Number of patients and images used for training, validating, and testing the network.

Training		**Validation**		**Testing**	
# of Patients	# of Images	# of Patients	# of Images	# of Patients	# of Images
61	682	3	85	5	137

Network Description

Segmentations were created using a supervised U-Net architecture constructed in MATLAB 2019a, shown in Figure 2[10]. We modified a pre-existing architecture to achieve our desired results. The original U-Net architecture consists of an encoder subnetwork and decoder subnetwork that are connected by a bridge section. The encoder and decoder subnetworks in the U-Net architecture consists of multiple stages. EncoderDepth, which specifies the depth of the encoder and decoder subnetworks, sets the number of stages. The stages within the U-Net encoder subnetwork consist of two sets of convolutional and rectified linear unit (ReLU) layers, followed by a 2x2 max pooling layer. The decoder subnetwork consists of a transposed convolution layer for upsampling, followed by two sets of convolutional and ReLU layers. The bridge section consists of two sets of convolution and ReLU layers. The bias term of all convolutional layers is initialized to zero. Convolution layer weights in the encoder and decoder subnetworks are initialized using the method developed by He *et al*.[11].

Root mean squared propagation (RMSProp) was used as the optimizer. RMSProp adapts the learning rate with a moving average of the parameter gradients. Dice Similarity Coefficient was used as the classification loss function[12]. Training was performed for 25 epochs, with an initial learning rate of 10^{-4}, which was decreased by a factor of 0.25 every 4 epochs. These parameters were found empirically based on the DSC performance in the validation group when compared to the results from the training data. After segmentation by the network, objects less than 200 pixels in area are removed from each CT slice. This was also determined empirically using the training and validation segmentation results. Once the final model and post-processing methods were determined, the test data was evaluated once, to determine the generalizability of the method.

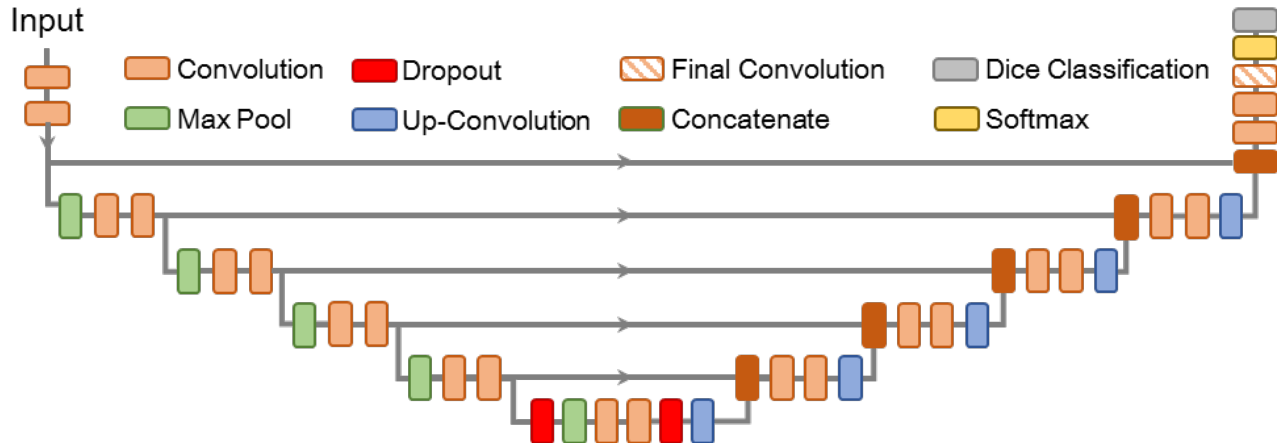

Figure 2. U-Net architecture used for the segmentation.

Evaluation Metrics

We evaluated the robustness of the trained CNN by calculating the following metrics: Dice Similarity Coefficient (DSC), precision, recall[13], and absolute percent area change from ground truth in muscle between the reference and network segmentation. DSC *(Equation 1)* describes the degree of overlap between a ground truth segmentation ***A*** and the segmentation ***B*** produced by the CNN. A value of 1 indicates perfect overlap between the two images.

$$\text{Dice Similarity Coeffiecient} = \frac{2(\boldsymbol{A} \cap \boldsymbol{B})}{(\boldsymbol{A} + \boldsymbol{B})} \quad (1)$$

Recall *(Equation 2)* quantifies how accurately the network performed when identifying the abdominal muscle in the image. Here, true positives are muscle pixels correctly classified as muscle by the network, while false negatives refer to background pixels incorrectly classified as muscle. A recall value of 1 indicates all pixels labeled by the network as muscle are in fact muscle.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

Precision *(Equation 3)* is a measurement of the amount of muscle the network detected in the image. In addition to true positives described above, precision also uses the number of false positives, or pixels of muscle that were classified as background by the network.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

Absolute percent area difference from ground truth Δ_{GT} *(Equation 4)* is a metric designed to make a quantitative characterization of the robustness of the network by comparing the area established in the ground truth segmentations made by the specialist, and the segmentations created by the CNN. This is useful to evaluate the method for clinical applications in which the end goal to monitor the change of muscle mass of a patient over time.

$$\Delta_{GT} = \left| \frac{Area\ of\ Input - Area\ of\ Output}{Area\ of\ Input} \times 100 \right| \quad (4)$$

RESULTS

Results for the evaluation metrics for the network segmentations are shown in Table II. All metrics had low variability, therefore the standard deviations are very narrow. This suggests the CNN model had robust performance across all images. The mean DSC for the testing group was 0.92, with a standard deviation of 0.024. This result indicates the ability for this CNN to obtain segmentations very similar to that of the trained specialist. Mean recall and precision values for the test data were 0.91 +/- 0.036 and 0.93 +/- 0.033, respectively; meaning that our network was able to correctly segment abdominal muscle for the majority of the independent test data cases presented. The absolute percent area difference obtained from the testing data was 5.0 +/- 0.034. This metric is comparable to the absolute percent area difference obtained for the training data, and an improvement from the validation data. This demonstrates that overall, the CNN correctly segmented the area associated with abdominal muscle 95% of the time.

Evaluation metric distributions in the test dataset are shown as individual histograms in Figure 3. Segmentation results for the worst case, average case, and best case are displayed in Figure 4, respectively. We labeled these cases by comparing DSC values of the testing data. We chose DSC to make this assessment because obtaining segmentations with high similarity to the segmentations made by the physician is arguably the most important goal of this work.

Table II. Average +/- Standard Deviation for Segmentation Results per Image.

	DSC	Recall	Precision	|% Area Diff.|
Training	0.92 +/- 0.032	0.92 +/- 0.035	0.91 +/- 0.046	4.6 +/- 0.045
Validation	0.92 +/- 0.035	0.94 +/- 0.021	0.89 +/- 0.057	6.3 +/- 0.069
Testing	0.92 +/- 0.024	0.91 +/- 0.036	0.93 +/- 0.033	5.0 +/- 0.034

The histograms shown below reiterates the results displayed in Table II above. For the test data DSC, the values ranged from 0.86 to 0.95, demonstrating consistent network performance. For recall and precision, the results ranged from 0.81 to 0.96, and 0.82 to 0.98, respectively. This suggests no noticeable bias in the network to over segmentation or undersegmentation. The percent area difference from ground truth varied by up to 16% in the worst case. However, 59% of images had a Δ_{GT} below 5% and 89% of images had a Δ_{GT} below 10%.

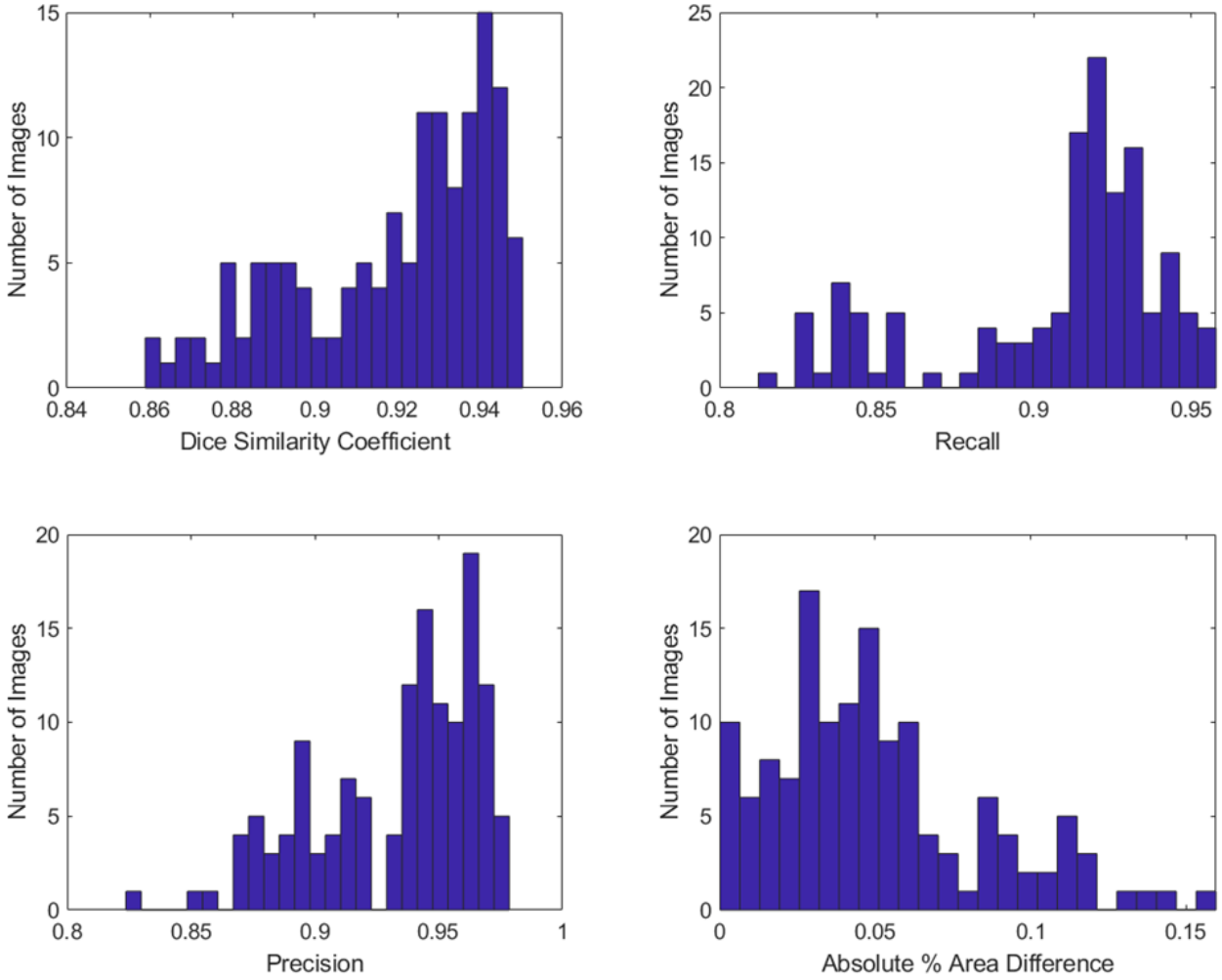

Figure 3: Histogram of the DSC, recall, precision, and absolute percent area difference values from the test dataset (N = 137 images).

Qualitative results from the segmentation for select images are shown in Figure 4, where the dark blue background represents the original CT image and the light blue is the segmentation generated by the CNN. In Figure 4 A1-A3, the three worst segmentation results are shown, with DSC values of 0.86 to 0.87. For Figure 4 A1, both oversegmentation and undersegmentation are observed, while for A2 and A3, undersegmentation is the predominate cause of error. Figure 4 B1-B3 show average segmentation results from our CNN, each with a DSC of 0.92, with slight undersegmentation seen in B1 and slight oversegmentation observed in B2 and B3. Figure 4 C1-C3 represent the best results, each with a DSC of 0.95. Minute oversegmentation is seen in C2. Overall, the most common source of oversegmentation was seen involving the liver, suggesting the network had difficulty identifying the tissue barrier between the liver and the abdominal muscle is some cases.

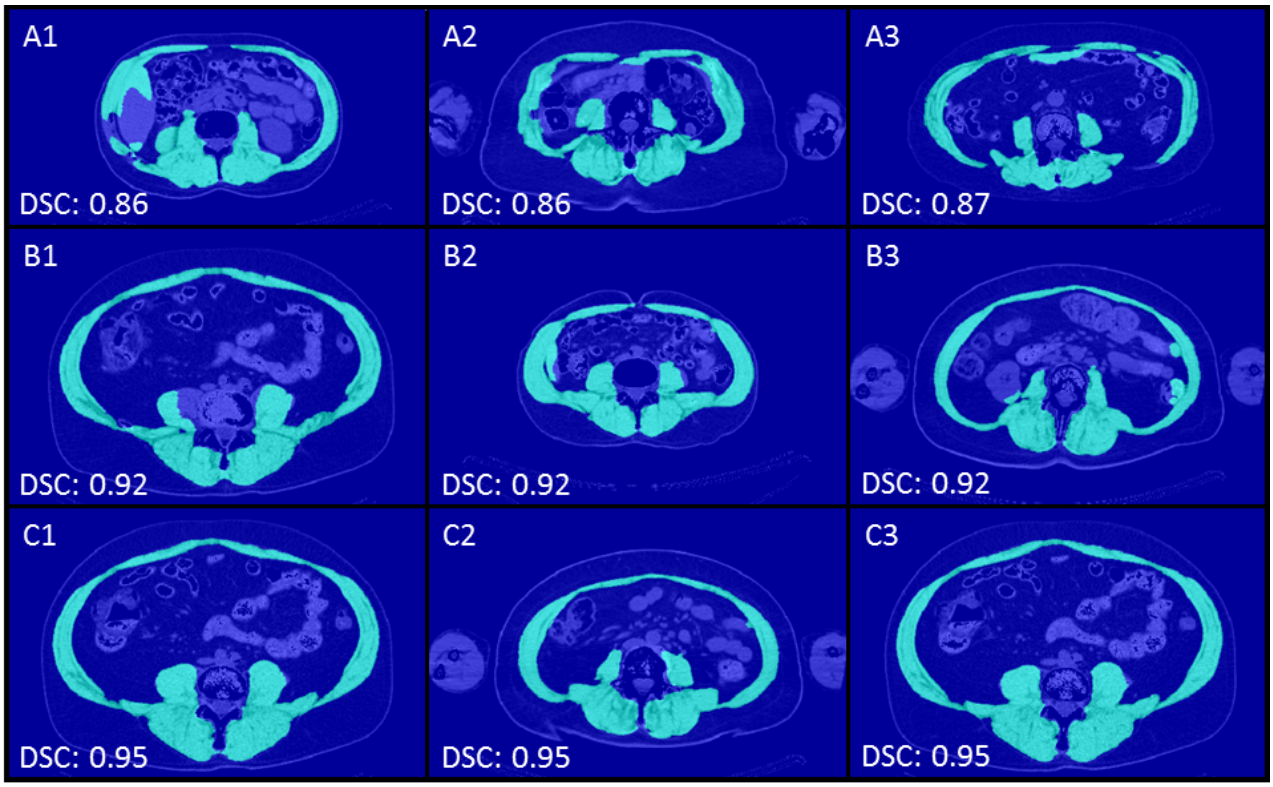

Figure 4: Examples of CT images with low (A1-A3), average (B1-B3), and high (C1-C3) DSC results. Among the images, both oversegmentation and undersegmentation are observed.

DISCUSSION

The consistent DSC, precision, recall, and percent area difference from ground truth values between the groups (Table II & Figure 3) suggests the neural network model is well trained without being over-fit to either the training data or the validation data. In particular, the similar values for precision and recall suggest the network has no bias towards oversegmentation or undersegmentation. The mean vales for DSC, precision, and recall each exceeds 90%, which supports the ability for our CNN model to effectively segment skeletal abdominal muscle comparably to a trained specialist.

The most common oversegmentation occurred in regions near the liver, as seen in Figure 4 A1. This is probably due to the similar textures present in the abdominal muscle and liver muscle as represented in the CT images or difficulty in determining organ boundaries between the liver and abdominal muscle. Additionally, several patients with arms in the imaging field had portions of the forearm segmented. The presence of arms is caused by inconsistent CT image acquisition in the clinical setting. As a result of our limited training data, our CNN was unable to classify the arms in some cases.

In addition to oversegmentation, undersegmentation occurred in some narrow regions of abdominal muscle, most notably seen in Figure 4 A2 and A3. This could be due to the presence of fat between muscles or fatty-replaced muscles, which would have been segmented in the reference but might be missed by the neural network as the texture would appear different when compared to the neighboring muscle. Anatomical differences, such as intramuscular fat, vary highly between individuals and presents itself as a limitation of the deep CNN solution without extensive training. This further proves the importance of introducing many different patients to ensure as much variability as possible in the

training data set. In cases such as these, the evaluation metrics for the network performance would improve if the reference standard accounted for the small regions of fat.

CONCLUSION

A U-Net was used to segment abdominal muscle of patients by 2D image analysis, with a mean DSC value of 0.92 for an independent test dataset. These segmentations had a mean absolute percent area difference of 5.0% when compared to the reference standard. These segmentations could be useful to physicians who desire to track changes in abdominal muscle throughout treatment in order to determine how a patient is responding. By automating the segmentation process, the physician can spend less time performing tedious tasks such as manual segmentation, and more time with the patient.

In this work, we present an accurate segmentation of abdominal muscle on CT slices using U-Net and simple pre- and post-processing methods. The segmentations are accurate in terms of both DSC and absolute percent area difference, suggesting that the method could be useful as an automated way of tracking changes in abdominal muscle over time. This method can also be used as a possible indication of how a patient is responding during treatment. Additionally, several patients with arms in the imaging field had portions of the forearm segmented. This could be improved by implementing another classification that delineates background, abdominal muscle, and other muscle. This solution would require several abdominal CT images that include arm segmentations as well.

Another issue we faced was the undersegmentation which occurred in narrow regions of abdominal muscle, as seen in Figure 4 A2 and A3. This could be corrected by comparing neighboring slices from the results of the neural network. If there is a significant difference between two slices, the CNN may have incorrectly labeled an area. Post-processing can be implemented to check for missing abdominal muscle. Then, we could fill any gaps created by undersegmentation. The limitation to this approach is undermining significant muscle mass changes that may be characteristic of sarcopenia. Further understanding of what determines a “significant” skeletal abdominal muscle mass changes must be understood further to introduce post-processing image correction in the clinical setting.

Further work to improve the robustness of our neural network will increase the efficacy of our model for segmenting CT images. We used DSC, precision and recall to quantify the efficacy of our model. We achieved a mean DSC of 0.92, mean precision of 0.93, and mean recall of 0.91 in an independent test set. Percentage area difference from ground truth was also quantified in this study, with a mean value of 5% in the testing data. Also, volume analysis could also support the capabilities of our automatic segmentation CNN model. We can use our current model to analyze percentage volume area difference by catenating our 2D images into 3D images. This will allow the radiologist to gain more information about differences in volumetric muscle loss.

Acknowledgements

This research was supported in part by the U.S. National Institutes of Health (NIH) grants (R01CA156775, R01CA204254, R01HL140325, and R21CA231911) and by the Cancer Prevention and Research Institute of Texas (CPRIT) grant RP190588.

References

[1] Graffy, P., Liu, J., Pickhardt, P., Burns, J., Yao, J., Yao, J., and Summers, R., "Deep Learning-Based muscle segmentation and quantification at abdominal CT: Application to a longitudinal adult screening cohort for Sarcopenia assessment," The British journal of radiology, 92(xxxx), 20190327 (2019).

[2] Nishimura, J. M., Ansari, A. Z., D'Souza, D. M., Moffatt-Bruce, S. D., Merritt, R. E., and Kneuertz, P. J., "Computed Tomography-Assessed Skeletal Muscle Mass as a Predictor of Outcomes in Lung Cancer Surgery," The Annals of thoracic surgery, (2019).

[3] Xia, W., Barazanchi, A. W., MacFater, W. S., and Hill, A. G., "The impact of computed tomography-assessed sarcopenia on outcomes for trauma patients–a systematic review and meta-analysis," Injury, (2019).

[4] Kalyani, R. R., Corriere, M., and Ferrucci, L., "Age-related and disease-related muscle loss: the effect of diabetes, obesity, and other diseases," The lancet. Diabetes & endocrinology, 2(10), 819-829 (2014).

[5] Hemke, R., Buckless, C. G., Tsao, A., Wang, B., and Torriani, M., "Deep learning for automated segmentation of pelvic muscles, fat, and bone from CT studies for body composition assessment," Skeletal Radiology, 1-9 (2019).

[6] Wang, Y., Qiu, Y., Thai, T., Moore, K., Liu, H., and Zheng, B., "A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images," Computer methods and programs in biomedicine, 144, 97-104 (2017).

[7] Barnard, R., Tan, J., Roller, B., Chiles, C., Weaver, A. A., Boutin, R. D., Kritchevsky, S. B., and Lenchik, L., "Machine Learning for Automatic Paraspinous Muscle Area and Attenuation Measures on Low-Dose Chest CT Scans," Academic radiology, (2019).

[8] Weston, A. D., Korfiatis, P., Kline, T. L., Philbrick, K. A., Kostandy, P., Sakinis, T., Sugimoto, M., Takahashi, N., and Erickson, B. J., "Automated abdominal segmentation of CT scans for body composition analysis using deep learning," Radiology, 290(3), 669-679 (2018).

[9] Shahedi, M., Ma, L., Halicek, M., Guo, R., Zhang, G., Schuster, D. M., Nieh, P., Master, V., and Fei, B., [A semiautomatic algorithm for three-dimensional segmentation of the prostate on CT images using shape and local texture characteristics] SPIE, MI (2018).

[10] Ronneberger, O., Fischer, P., and Brox, T., [U-Net: Convolutional Networks for Biomedical Image Segmentation] Springer International Publishing, Cham(2015).

[11] He, K., Zhang, X., Ren, S., and Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." 1026-1034.

[12] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J., [Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations] Springer, (2017).

[13] Shahedi, M., Cool, D. W., Romagnoli, C., Bauman, G. S., Bastian-Jordan, M., Gibson, E., Rodrigues, G., Ahmad, B., Lock, M., and Fenster, A., "Spatially varying accuracy and reproducibility of prostate segmentation in magnetic resonance images using manual and semiautomated methods," Medical physics, 41(11), 113503 (2014).