Original Article | Gastrointestinal Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2019.0470
Korean J Radiol 2020;21(1):88-100

Korean Journal of Radiology
KJR

Development and Validation of a Deep Learning System for Segmentation of Abdominal Muscle and Fat on Computed Tomography

Hyo Jung Park, MD[1], Yongbin Shin, PhD[2], Jisuk Park, MS[1], Hyosang Kim, MD, PhD[3], In Seob Lee, MD, PhD[4], Dong-Woo Seo, MD, PhD[5], Jimi Huh, MD, PhD[6], Tae Young Lee, MD[7], TaeYong Park, PhD[2], Jeongjin Lee, PhD[2], Kyung Won Kim, MD, PhD[1]

[1]Department of Radiology and Research Institute of Radiology, Asan Image Metrics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea; [2]School of Computer Science and Engineering, Soongsil University, Seoul, Korea; Departments of [3]Nephrology, Internal Medicine, [4]Surgery, and [5]Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea; [6]Department of Radiology, Ajou University School of Medicine and Graduate School of Medicine, Ajou University Hospital, Suwon, Korea; [7]Department of Radiology, Ulsan University Hospital, Ulsan, Korea

Objective: We aimed to develop and validate a deep learning system for fully automated segmentation of abdominal muscle and fat areas on computed tomography (CT) images.

Materials and Methods: A fully convolutional network-based segmentation system was developed using a training dataset of 883 CT scans from 467 subjects. Axial CT images obtained at the inferior endplate level of the 3rd lumbar vertebra were used for the analysis. Manually drawn segmentation maps of the skeletal muscle, visceral fat, and subcutaneous fat were created to serve as ground truth data. The performance of the fully convolutional network-based segmentation system was evaluated using the Dice similarity coefficient and cross-sectional area error, for both a separate internal validation dataset (426 CT scans from 308 subjects) and an external validation dataset (171 CT scans from 171 subjects from two outside hospitals).

Results: The mean Dice similarity coefficients for muscle, subcutaneous fat, and visceral fat were high for both the internal (0.96, 0.97, and 0.97, respectively) and external (0.97, 0.97, and 0.97, respectively) validation datasets, while the mean cross-sectional area errors for muscle, subcutaneous fat, and visceral fat were low for both internal (2.1%, 3.8%, and 1.8%, respectively) and external (2.7%, 4.6%, and 2.3%, respectively) validation datasets.

Conclusion: The fully convolutional network-based segmentation system exhibited high performance and accuracy in the automatic segmentation of abdominal muscle and fat on CT images.

Keywords: *Deep learning; Artificial intelligence; Sarcopenia; Muscles; Adipose tissue*

INTRODUCTION

Quantification of the muscle and fat mass of the body, which is referred to as body morphometry, is an essential part of health evaluation. Increasing rates of occurrence of obesity have been observed in the past few decades, and extensive research has been conducted to explore the relationship between body fat and various diseases (1, 2). Recently, sarcopenia (the loss of skeletal muscle mass) has emerged as a biomarker for assessing the prognosis in various cancer patients (3-5), morbidity/mortality rate after major surgery (6-8), and general health status (9).

Received June 26, 2019; accepted after revision October 15, 2019.
This study was supported by a grant of the Korea Health Industry Development Institute (No. HI18C1216) and a grant of the National Research Foundation of Korea (No. 2017R1A2B3011475).

Corresponding author: Kyung Won Kim, MD, PhD, Department of Radiology and Research Institute of Radiology, Asan Image Metrics, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.
• Tel: (822) 3010-4377 • Fax: (822) 476-4719 • E-mail: medimash@gmail.com

Korean Journal of Radiology
KJR

CT is a useful tool for evaluating body muscle and fat, as areas of muscle and fat tissue can be accurately visualized on the basis of CT attenuation values. Recent evidence suggests that abdominal CT-based measurements of muscle mass correlate well with the actual muscle mass (10, 11), and abdominal CT has also been proven as an accurate method for quantitative assessment of abdominal visceral fat (12).

Semi-automated segmentation methods requiring manual error correction or complex hand-crafted, feature-selection processes have been used for the segmentation of muscle and fat on cross-sectional images (10, 13-16). However, these approaches cannot usually be applied to large datasets because of the human resource and time requirements. These limitations might be overcome by a fully automated segmentation method using deep learning. With its power to use neural networks and convolutional layers to learn the hierarchy of features in a large amount of given data (17), deep learning systems can be trained to analyze body morphometry. A deep learning system that could automatically and accurately segment muscle and fat on CT images would be useful for clinical practice and various research purposes requiring body morphometry analysis.

There have been several studies reporting the accuracy of automatic quantification of abdominal muscle area using deep learning (18-22). The algorithms used were different across studies, such as U-net neural network model, multi-atlas segmentation model, fully convolutional network (FCN), and augmented active shape model. In this study, we aimed to develop a deep learning system using the FCN model combined with imaging pre-processing for automated segmentation of muscle on abdominal CT images and to evaluate its performance on internal and external validation sets.

MATERIALS AND METHODS

Study Subjects

This study was approved by the Institutional Review Board (IRB) of the three participating institutions, and the requirement for informed consent was waived for this retrospective analysis. This study included CT examinations from subjects with gastric cancer, pancreatic cancer, and sepsis and healthy individuals who were referred to the imaging core lab of our institution by various physicians at the three participating institutions for body morphometric analysis. The subjects were consecutively recruited for separate clinical studies of the referring physicians (not yet published). Under the agreement of referring physicians and IRB approval, we included the abdominal CT scans referred to our imaging core lab in 2017 and 2018 for abdominal muscle quantification.

The deep learning system was developed using a training dataset of 883 CT scans from 467 subjects referred from our hospital in 2017. The trained system was then validated on two independent datasets of 479 total subjects: a temporally distinct internal validation dataset of 426 CT scans from 308 subjects referred from our hospital in 2018 and an external validation dataset of 171 CT scans from 171 subjects referred from two outside hospitals in 2018 (Fig. 1). The clinical characteristics of the subjects in the training and validation datasets are summarized in Table 1. In the datasets, the gastric and pancreatic cancer groups included patients who were surgically treated with curative intent, ranging from stage I to III according to the AJCC 8th edition. The sepsis group included patients who were admitted to the emergency department for septic shock with refractory hypotension or hypoperfusion. The no disease group included subjects undergoing liver donor work-up (training and internal validation sets) or routine check-up (external validation set). A subset of patients had two or more CT scans for evaluation of chronological change of body morphometry (i.e., before and after gastrectomy or treatment for sepsis). In our study, we used the CT scan as a basic unit for training and validation.

CT Imaging

CT scanners from various manufacturers (Sensation 16, Somatom Definition, Somatom Definition flash, and Somatom Definition AS + scanners, Siemens Healthineers, Erlangen, Germany; LightSpeed 16, LightSpeed plus, and LightSpeed VCT scanners, GE Healthcare, Milwaukee, WI, USA; Ingenuity and iCT 256 scanners, Philips Healthcare, Amsterdam, Netherlands; and Aquilion scanner, Toshiba, Tokyo, Japan) were used. In training and validation datasets as a whole, Siemens scanners were most commonly used (n = 974), followed by GE (n = 361), Philips (n = 111), and others (n = 34), as detailed in Supplementary Table 1. The scan parameters varied according to the institutional policy at each participating hospital; tube voltage was 80, 100, 120, or 140 kVp and tube current was 170–250 mAs. No low-dose CT protocol was included in the study datasets. The images were reconstructed in the axial plane and ranged from 2.5 to 5 mm in thickness at 2.5 to 5 mm intervals.

 https://doi.org/10.3348/kjr.2019.0470

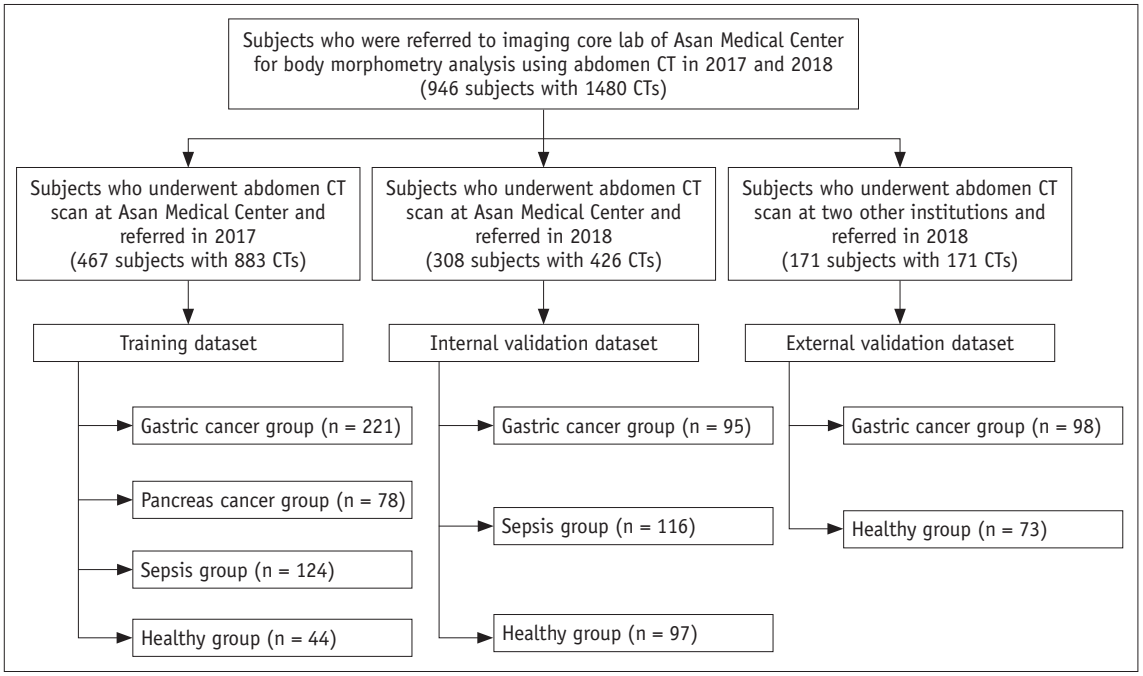


Fig. 1. Overview of patient recruitment process.

Table 1. Subject Characteristics of Training and Validation Datasets

Characteristics	Training Dataset	Internal Validation Dataset	External Validation Dataset
Number of subjects*	467 (883 CTs)	308 (426 CTs)	171 (171 CTs)
Age (years)†	56.1 ± 13.9 (18–86)	56.6 ± 14.2 (20–86)	61.1 ± 11.1 (18–85)
Female (%, female:male)	36.6 (171:296)	44.1 (136:172)	39.8 (68:103)
Height (m)†	1.7 ± 0.1 (1.4–1.9)	1.6 ± 0.1 (1.4–1.8)	1.6 ± 0.1 (1.4–1.8)
Weight (kg)†	62.1 ± 11.0 (36.0–97.2)	63.6 ± 12.2 (34.0–115.7)	65.2 ± 10.2 (36.8–95.4)
BMI (kg/m^2)†	23.2 ± 3.3 (13.9–37.7)	23.4 ± 3.5 (15.7–45.2)	24.6 ± 3.1 (18.0–36.5)
Underlying disease (n)*			
Gastric cancer	221 (497 CTs)	95 (95 CTs)	98 (98 CTs)
Sepsis	124 (264 CTs)	116 (234 CTs)	0
Pancreatic cancer	78 (78 CTs)	0	0
No disease	44 (44 CTs)	97 (97 CTs)	73 (73 CTs)

*Data are number of subjects, with number of CT scans in parenthesis, †Data are mean ± standard deviation, with range in parenthesis.
BMI = body mass index

Images were reconstructed using filtered back projection (B30f, B30s, B41f, B41s) or iterative reconstruction (I30s, I30f). Pixel size ranged from 0.57 to 0.77 mm.

The training and validation datasets contained both contrast-enhanced CT scans (977 scans) and non-enhanced CT scans (503 scans). The proportions of non-enhanced CT scans were 20.7% (183 scans), 56.5% (241 scans), and 46.2% (79 scans) for the training, internal validation, and external validation datasets, respectively. For the contrast-enhanced CT scans, contrast medium (100–150 mL of 300–370 mgI/mL non-ionic iodine contrast) was intravenously administered at a rate of 2–3 mL/sec and portal venous phase images were obtained by using a fixed 75-second delay.

Generation of Ground Truth Labels

A single axial image from the level of the inferior endplate of the 3rd lumbar vertebra was extracted from each CT scan and used for the analysis. Ground truth segmentation maps were created using semi-automated segmentation software (AsanJ-Morphometry™, Asan Image Metrics, Seoul, Korea, http://datasharing.aim-aicro.com/morphometry). An expert image analyst (with 8 years' experience) created segmentation maps of abdominal wall muscle, visceral fat, and subcutaneous fat, which served as ground truth

https://doi.org/10.3348/kjr.2019.0470

Korean Journal of Radiology
KJR

labels. The abdominal muscles included psoas, paraspinal, transversus abdominis, rectus abdominis, quadratus lumborum, and internal and external obliques and were demarcated using predetermined thresholds (-29 to +150 Hounsfield unit [HU]) (23). The visceral and superficial fat was also demarcated using fat tissue thresholds (-190 to -30 HU) (23). The total number of pixels and cross-sectional area (CSA; cm^2) of the segmented regions were obtained. All segmentation maps were reviewed and corrected as necessary by a board-certified abdominal radiologist (with 5 years of clinical experience).

Development of the FCN-Based Segmentation System

The FCN-based segmentation system was developed using supervised learning of the ground truth labels (24). By using data augmentation, 11167 training data images were generated from the 883 CT scans. The data augmentation was performed using random combinations of affine transformations, Gaussian filtering, and anisotropic diffusion filtering (25-27).

To generate the input data for training our FCN-based segmentation system, the HU information required conversion to grayscale information. However, converting each CT image to grayscale with the commonly used HU range for the target tissue might have yielded inaccurate segmentation results, as the HU values vary according to the scanner type and scanning protocol (28). Several pre-processing steps were used to overcome this limitation: Otsu thresholding, seeded region growing, morphological filtering, and histogram equalization (29, 30). First, using Otsu thresholding, a region of a patient's image was extracted from the background. Then, hole filling and noise removal were performed on the region by seeded region growing and morphological filtering. Finally, the histogram distribution of the region was normalized to generate consistent grayscale information, irrespective of the scanner type and protocol.

In deep learning-based imaging processes, the image resolution is lowered by the application of multiple convolutional layers and pooling processes (17); this is generally not problematic for image detection or classification tasks. However, for segmentation of a certain region in an image, adequate image resolution is key to obtaining accurate results. By adding upsampling layers, the FCN enables pixel-wise segmentation of each body region by producing output layers with a resolution restored to that of the original input image (24). Furthermore, more refined results can be obtained by fusing information from layers with different strides; to this end, the fusion process was extended by fusing the final layer and pooling layers 4 to 1 (Fig. 2). During training, the segmentation results were assessed three times, both qualitatively and quantitatively, by two board-certified abdominal radiologists, and recognized errors were back-propagated to optimize the architecture and parameters of the neural network. All experiments were run on Intel® Core™ i7–7700K GPU (8M Cache, 4.20 GHz, Santa Clara, CA, USA).

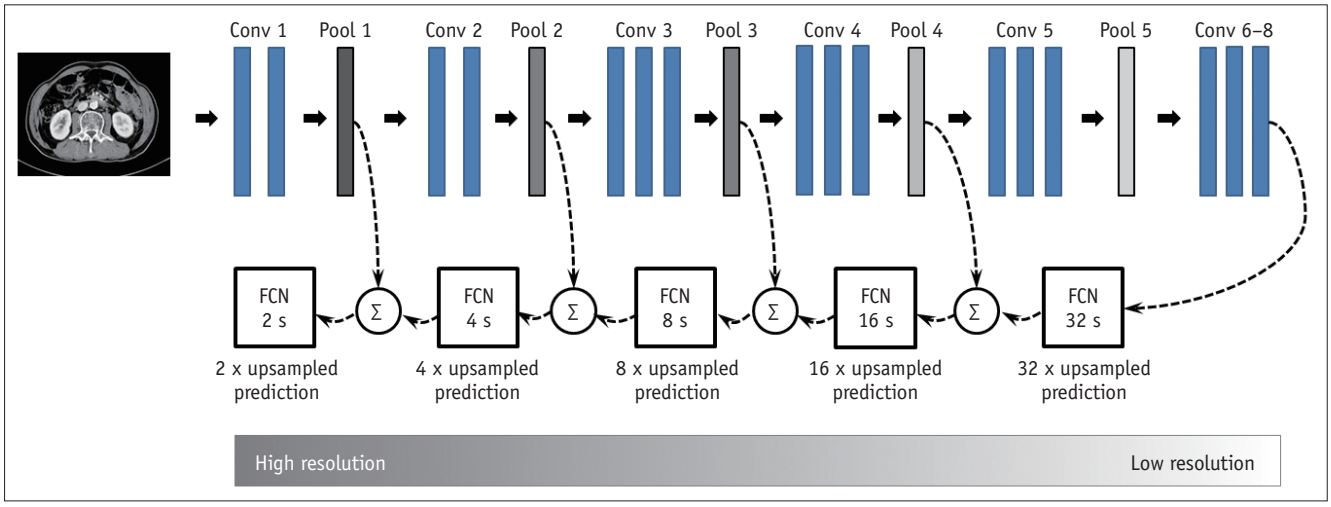


Fig. 2. Overview of FCN. In our FCN training process, several upsampling layers were added, which enabled convolutional network to produce output layers with image resolution restored to original dimensions. FCN-32 s up-samples stride 32 predictions back to pixels. FCN-16 s combines predictions from both final layer and pooling 4 layers, allowing net to predict finer details while retaining high-level semantic information. FCN-8 s, FCN-4 s, and FCN-2 s receive additional predictions from pooling 3, pooling 2, and pooling 1, respectively, and thereby provide further precision. Conv = convolutional layer, FCN = fully convolutional network

https://doi.org/10.3348/kjr.2019.0470

Clinical Validation

Segmentation Performance of the Deep Learning System

The accuracy of the FCN-based segmentation was validated with separate internal and external validation datasets. After data of validation datasets were input into the system, it returned results in the form of the total number of pixels and CSA (cm^2) of skeletal muscle, subcutaneous fat, and visceral fat. An overview of the segmentation process using training and validation datasets is depicted in Figure 3.

The performance of the FCN-based segmentation system was evaluated using the Dice similarity coefficient (DSC) and CSA error. The DSC is a spatial overlap index and reproducibility validation metric ranging from 0 (no spatial overlap) to 1 (complete overlap) (31). The DSC was used to compare the degree of overlap between the number of pixels of the ground truth data and the deep learning system-derived segmentation data, which was calculated according to the following equation:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where true positive (TP) represents the number of pixels correctly detected (i.e., pixels included in both ground truth labels and FCN-based segmentation), false positive (FP) represents the number of pixels falsely detected (i.e., pixels included in the FCN-based segmentation but not in the ground truth labels), and false negative (FN) represents the number of pixels included in the ground truth labels but not in the FCN-based segmentation. Thus, a high DSC implies high segmentation performance.

CSA error represents a standardized measure of the percentage difference in the area between the ground truth labels and FCN-based segmentation; thus, a low CSA error implies high segmentation performance, and was calculated as follows:

$$\text{CSA error (\%)} = \frac{|\text{Ground truth}_{CSA} - \text{FCN-based segmentation}_{CSA}|}{\text{Ground truth}_{CSA}} \times 100$$

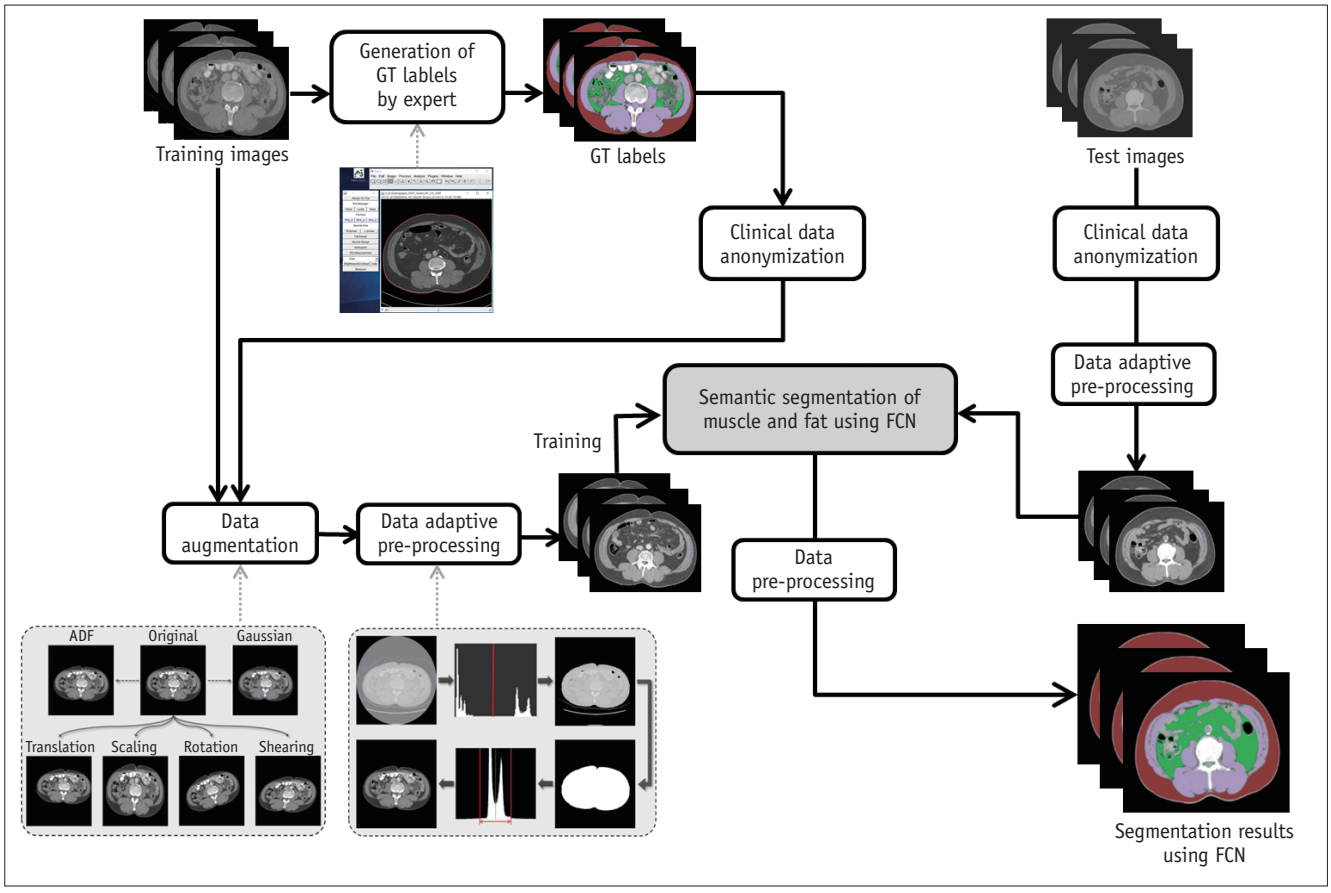


Fig. 3. Overview of FCN-based segmentation system. ADF = anisotropic diffusion filter, GT = ground truth

https://doi.org/10.3348/kjr.2019.0470

Korean Journal of Radiology
KJR

In addition, to evaluate the spatial accuracy of the FCN segmentation results (i.e., whether each region was segmented in the correct location), two board-certified abdominal radiologists performed visual assessment of all FCN-based segmentation maps from the validation datasets. Images were classified as spatially accurate or requiring manual adjustment, and the region requiring adjustment was identified.

Factors Influencing the Performance of the Deep Learning System

To explore any factors influencing the segmentation performance of the FCN-based segmentation system, subgroup analyses were performed according to intravenous contrast use, size of the segmented area, and disease type of the patients in the validation datasets. To evaluate the influence of the intravenous contrast, the segmentation performance was compared between CT scans acquired with (46.4%, 277/597) and without (53.6%, 320/597) intravenous contrast media in all validation datasets. To evaluate the effect of the segmented area, a total of 597 CT scans from all validation datasets were divided into subgroups according to the quartiles of muscle area, subcutaneous fat area, and visceral fat area. The relationship of body mass index and segmentation performance was also investigated. Lastly, the influence of disease type was evaluated only in the internal validation dataset, for which clinical information was available. A total of 426 CT scans were divided according to disease type into a gastric cancer group (n = 95), sepsis group (n = 234), or healthy group (n = 97), and the performance of the FCN-based segmentation system was compared across the subgroups.

Statistical Analysis

Agreement of the segmentation results between the ground truth labels and FCN-based segmentation was measured using Bland-Altman plots. The limits of agreement of the Bland-Altman plots were defined as the mean differences ± 95% confidence intervals. To explore any factors influencing the performance of the FCN-based segmentation system, pairwise comparisons between subgroups were performed using independent *t* tests or analysis of variance with post-hoc *t* tests using Bonferroni correction. The correlation between body mass index and performance parameters was analyzed by Pearson correlation tests. SPSS version 21 (IBM Corp., Armonk, NY, USA) and MedCalc 12.7.0 (MedCalc Software, Mariakerke, Belgium) were used for statistical analyses. A p value < 0.05 was considered to represent statistical significance.

RESULTS

Performance of the FCN-Based Segmentation System

Our FCN-based segmentation system successfully generated segmentation results for all CT scans in the internal and external validation datasets. Segmentation process per image took 0.05 seconds on average. The results for all validation datasets (n = 597) are summarized in Table 2. The

Table 2. Performance of FCN-Based Segmentation System on Validation Datasets

Parameter	All Validation Datasets (n* = 597)			Internal Validation Dataset (n* = 426)			External Validation Dataset (n* = 171)		
	MA	SFA	VFA	MA	SFA	VFA	MA	SFA	VFA
CSA from ground truth (cm²)†	109.44 ± 31.86	134.27 ± 69.75	107.61 ± 56.26	107.78 ± 33.49	126.41 ± 67.10	104.77 ± 57.94	113.58 ± 27.02	153.86 ± 72.52	114.68 ± 51.33
CSA from FCN-based segmentation (cm²)†	110.11 ± 31.93	132.61 ± 68.33	106.55 ± 55.82	107.74 ± 33.21	126.63 ± 66.41	104.21 ± 57.69	116.01 ± 27.72	147.51 ± 70.93	112.40 ± 50.56
p value	0.715	0.677	0.744	0.987	0.962	0.886	0.411	0.413	0.679
DSC†	0.96 ± 0.02 (0.75–0.99)	0.97 ± 0.03 (0.34–0.99)	0.97 ± 0.01 (0.90–0.99)	0.96 ± 0.03 (0.75–0.99)	0.97 ± 0.04 (0.34–0.99)	0.97 ± 0.01 (0.90–0.99)	0.97 ± 0.01 (0.93–0.97)	0.97 ± 0.01 (0.93–0.99)	0.97 ± 0.01 (0.92–0.99)
CSA error (%)†	2.26 ± 2.41 (< 0.01–0.36)	4.02 ± 13.60 (0.03–326.36)	1.93 ± 1.80 (0.01–14.84)	2.08 ± 2.52 (< 0.01,–20.36)	3.79 ± 16.02 (0.03–326.36)	1.76 ± 1.83 (0.01–14.84)	2.70 ± 2.04 (< 0.01–10.51)	4.59 ± 2.54 (0.07–11.89)	2.33 ± 1.61 (0.01–8.60)

Data are presented as mean ± standard deviation, with range in parenthesis. *'n' indicates number of CT scans, †Data are mean ± standard deviation, with range in parenthesis. CSA = cross-sectional area, DSC = Dice similarity coefficient, FCN = fully convolutional network, MA = muscle area, SFA = subcutaneous fat area, VFA = visceral fat area

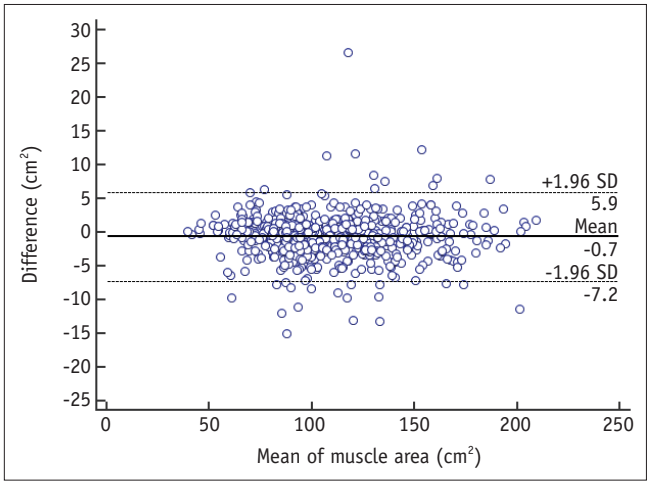
https://doi.org/10.3348/kjr.2019.0470

mean segmentation areas of muscle, subcutaneous fat, and visceral fat did not differ significantly between ground truth results and FCN-based segmentation results, for either the internal validation cohort (107.78 vs. 107.74 cm^2, 126.41 vs. 126.63 cm^2, and 104.77 vs. 104.21 cm^2, respectively; $p > 0.88$) or the external validation cohort (113.58 vs. 116.01 cm^2, 153.86 vs. 147.51 cm^2, and 114.68 vs. 112.40 cm^2, respectively; $p \geq 0.411$).

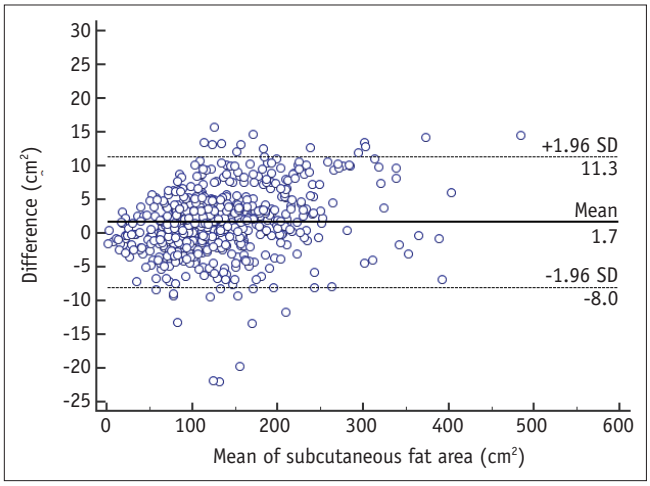
The mean DSC values for muscle, subcutaneous fat, and visceral fat were high for both the internal (0.96, 0.97, and 0.97, respectively) and external (0.97, 0.97, and 0.97, respectively) validation datasets. The mean CSA errors for muscle, subcutaneous fat, and visceral fat were low for both the internal (2.08%, 3.79%, and 1.76%, respectively) and external (2.70%, 4.59%, and 2.33%, respectively) validation datasets. One CT scan of a patient diagnosed with sepsis showed an exceptionally low DSC value (0.34) and high CSA error (326.36%) for subcutaneous fat; this case also showed the lowest CSA on ground truth and FCN-based segmentation (0.48 and 1.59 cm^2, respectively).

The Bland-Altman plots for all validation datasets (Fig. 4) indicated good agreement between ground truth measurements and those of the FCN-based segmentations for muscle, subcutaneous fat, and visceral fat, with mean differences (± limits of agreement) of 0.7 ± 6.6, 1.7 ± 9.3, and 1.1 ± 4.6 cm^2, respectively. Bland-Altman plots of the internal and external validation datasets are shown in Supplementary Figure 1.

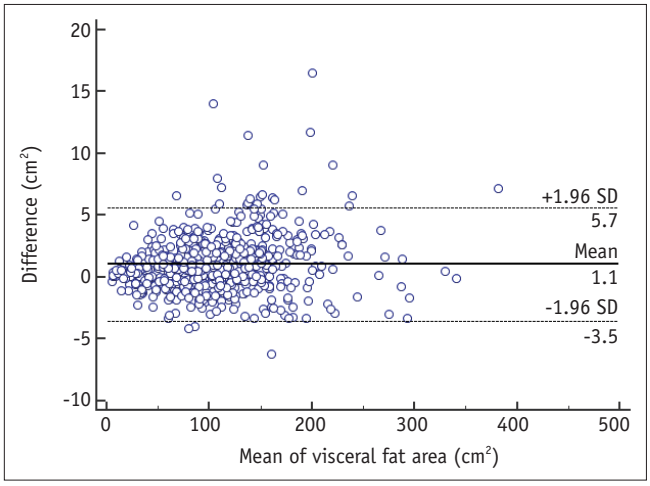
Assessment of the visual quality of the FCN-based segmentation maps revealed that 574 out of 597 CT scans (96.1%) were successfully segmented without requiring adjustment. Among the 23 CT scans requiring manual adjustment, 20 belonged to the sepsis group of the internal validation dataset, while the other three belonged to the external validation dataset and had metallic orthopedic hardware. Subcutaneous edema frequently (n = 18, 78.3%) caused under-segmentation of subcutaneous fat and over-segmentation of muscle. Metallic artifacts caused under-segmentation of adjacent tissue (n = 3, 13.0%). There was also a case of over-segmentation of muscle due to an enterostomy stump (n = 1, 4.3%), and a case of under-segmentation of muscle due to marked intramuscular fat infiltration (n = 1, 4.3%). The mean DSC values for muscle area, subcutaneous fat area, and visceral fat area of the 23 CT scans requiring manual adjustment were 0.92, 0.91, and 0.97, respectively, and the CSA errors of muscle area, subcutaneous fat area, and visceral fat area were 3.89%,



A



B



C

Fig. 4. Bland-Altman plots of muscle (A), subcutaneous fat (B), and visceral fat (C) for validation datasets.
Mean differences are equal to or less than 1.7 cm^2, with limits of agreement being equal to or less than 9.7 cm^2, suggesting comparable segmentation performance between ground truth and FCN-based segmentation results. SD = standard deviation

https://doi.org/10.3348/kjr.2019.0470

Korean Journal of Radiology
KJR

19.86%, and 2.71%, respectively (Supplementary Table 2). Representative examples of FCN-based segmentation are demonstrated in Figures 5 and 6.

Factors Influencing the Performance of the Deep Learning System

An overview of the subgroup analyses of the factors influencing the performance of the FCN-based segmentation system is presented in Table 3. The segmentation results for each subgroup are shown in Supplementary Table 3. In terms of the influence of intravenous contrast injection, the DSC values and CSA errors did not differ significantly between the contrast-enhanced group (n = 277) and non-enhanced group (n = 320) in any of the compared pairings ($p \geq 0.128$), indicating that intravenous contrast media had little effect on the performance of the FCN-based segmentation system.

Regarding the influence of the size of the segmented area, the group in the highest quartile of segmentation area (n = 149) had significantly higher DSC values and lower CSA errors than did the lowest quartile group ($p \leq 0.022$ in all pairs of comparison) for segmentation of muscle, subcutaneous fat, and visceral fat. Among the 20 CT scans in the sepsis group requiring manual adjustment of the FCN-based segmentation maps, 10 were in the lowest quartile group for subcutaneous fat. In particular, three CT scans from cachexic patients with a subcutaneous fat area less than 10 cm^2 (therefore included in the lowest quartile group for subcutaneous fat) showed high CSA errors of 20.18–326.36%. There were no significant correlations between body mass index and performance parameters (Supplementary Fig. 2).

As to the influence of the various diseases, of the three subgroups (gastric cancer group, sepsis group, and healthy group) in the internal validation dataset, the sepsis group showed the lowest DSC values and highest CSA errors in all areas of analysis. This group showed significantly smaller DSC values than did the other groups for muscle and

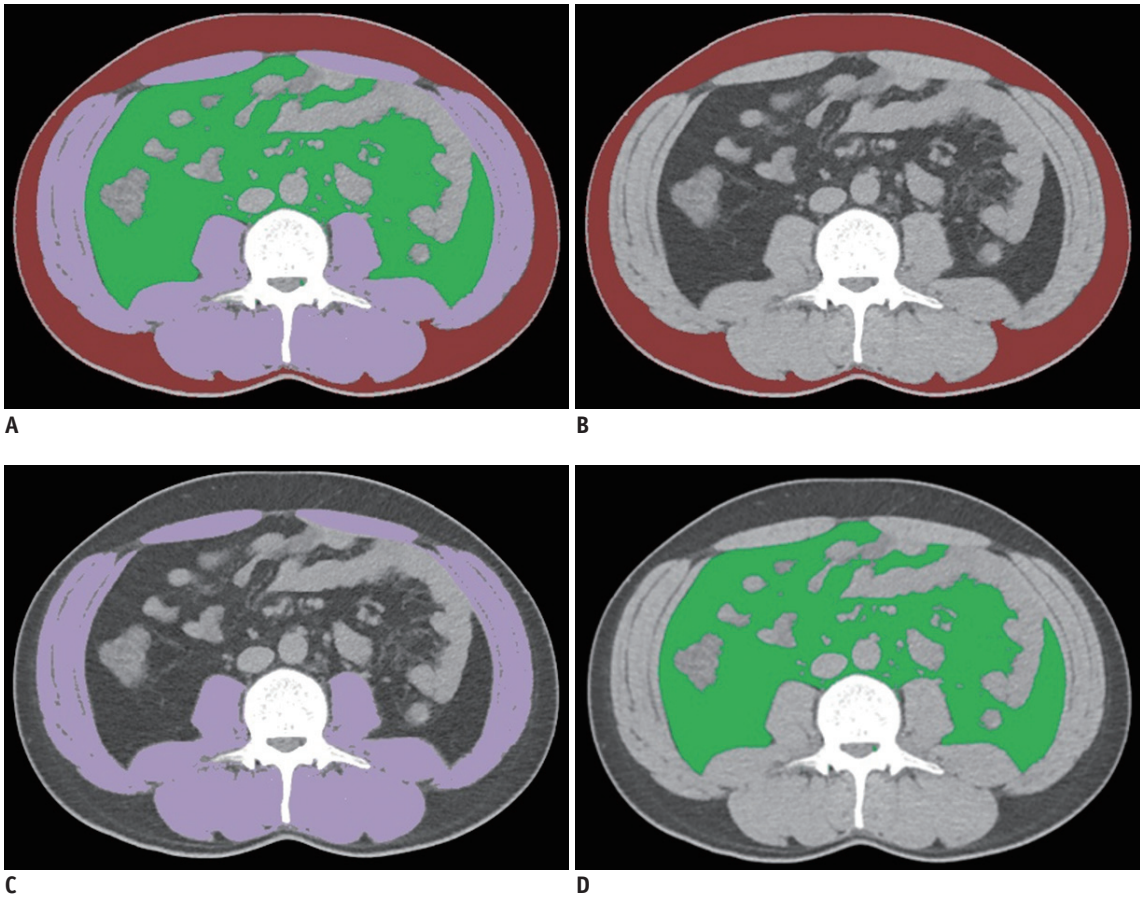


Fig. 5. Example of appropriately evaluated FCN-based segmentation map.
A. Fusion image of all segmented areas. **B-D.** Segmentation maps of subcutaneous fat (**B,** coded in red), skeletal muscle (**C,** coded in purple), and visceral fat (**D,** coded in green). Dice similarity coefficients are 0.98, 0.99, and 0.98 for subcutaneous fat, skeletal muscle, and visceral fat, respectively.

https://doi.org/10.3348/kjr.2019.0470

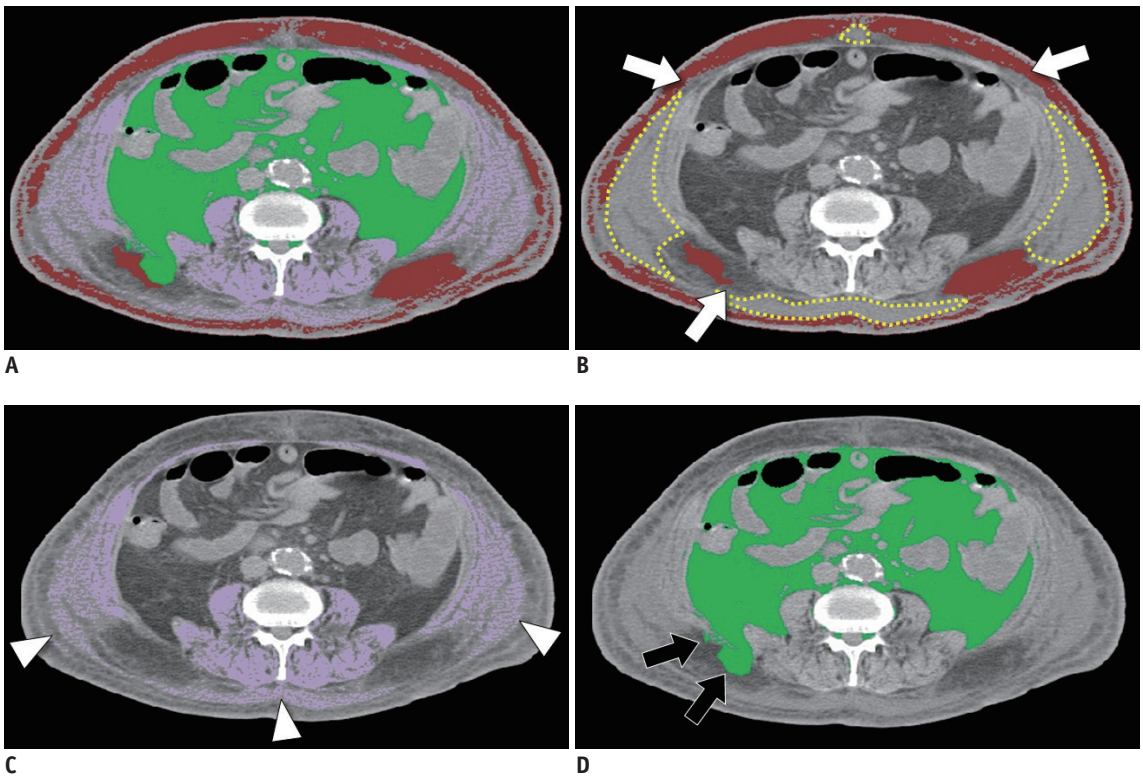


Fig. 6. Example of segmentation error.
A. Fusion image of all segmented areas. **B.** Segmentation map of subcutaneous fat (coded in red). There are areas with higher density compared with fat in subcutaneous area, which represent edema (dotted yellow line). Parts of subcutaneous fat abutting edema are not included in segmented subcutaneous fat (arrows). **C.** Segmentation map of skeletal muscle (coded in purple). Note subcutaneous edema segmented as muscle (arrowheads). **D.** Segmentation map of visceral fat. Some subcutaneous fat is erroneously segmented as visceral fat (arrows). Dice similarity coefficients are 0.78, 0.92, and 0.96 for subcutaneous fat, skeletal muscle, and visceral fat, respectively.

subcutaneous fat ($p < 0.001$), but not for visceral fat ($p = 0.644$). The sepsis group also showed significantly higher CSA error for muscle and visceral fat than did the other groups ($p < 0.001$), but non-significantly higher CSA error for subcutaneous fat and visceral fat ($p \geq 0.318$). Furthermore, the ground truth segmentation values of all areas were lowest in the sepsis group (Supplementary Table 3).

DISCUSSION

Our FCN-based segmentation system allowed accurate segmentation of muscle, subcutaneous fat, and visceral fat on abdominal CT images, showing DSC values higher than 0.96 and CSA errors lower than 5% for segmentations on the validation datasets. The differences between measurements by human experts and those by the FCN-based segmentation system were small (less than 10 cm^2), indicating comparable measurement performance of the FCN-based segmentation system. Segmentation maps created by the system were regarded as visually appropriate for 96.1% of the validation datasets. Our FCN-based segmentation system showed robust performance across different CT scanners and image acquisition protocols, for data from both our institution and from two outside institutions. In addition, it was not significantly influenced by intravenous contrast media injection. However, the performance of the FCN-based segmentation system could be influenced by various clinical conditions, including 1) conditions causing severe muscle/fat wasting (resulting in a low segmented area) such as cachexia and sarcopenia; 2) conditions causing fluid retention such as subcutaneous edema or ascites; and 3) conditions changing the normal anatomy such as abdominal surgery or spine surgery.

Our subgroup analysis demonstrated that the lowest quartile group showed the lowest segmentation performance for muscle, subcutaneous fat, and visceral fat. In particular, the segmentation error markedly increased in subjects with extremely small target segmentation areas. Almost half of the subjects in validation datasets requiring manual adjustment of the FCN-based segmentation maps belonged to the lowest quartile group for subcutaneous fat, including extremely cachexic patients with a subcutaneous fat area

https://doi.org/10.3348/kjr.2019.0470

Korean Journal of Radiology
KJR

Table 3. Subgroup Analyses to Evaluate Factors Influencing Performance of FCN-Based Segmentation System on Validation Cohorts

Influential Factors	Subgroups	DSC			CSA Error (%)		
		MA	SFA	VFA	MA	SFA	VFA
Contrast enhancement (n* = 597)	Used (n = 277)	0.96 ± 0.02 (0.84–0.99)	0.97 ± 0.02 (0.84–0.99)	0.97 ± 0.01 (0.92–0.99)	2.38 ± 1.96 (0.02–10.51)	3.35 ± 3.14 (0.05–23.30)	2.00 ± 1.92 (0.02–14.84)
	Not used (n = 320)	0.96 ± 0.00 (0.75–0.99)	0.97 ± 0.05 (0.34–0.99)	0.97 ± 0.03 (0.90–0.99)	2.15 ± 2.73 (< 0.01–20.36)	4.60 ± 18.34 (0.03–326.36)	1.86 ± 1.69 (0.01–12.74)
	p value†	0.927	0.128	0.369	0.231	0.229	0.350
Segmented area size (n* = 597)	Highest quartile (n = 149)	0.97 ± 0.02 (0.79–0.99)	0.98 ± 0.02 (0.87–0.99)	0.98 ± 0.00 (0.94–0.99)	1.80 ± 2.11 (0.03–20.36)	2.35 ± 1.67 (0.04–8.30)	1.61 ± 1.40 (0.01–8.01)
	Lowest quartile (n = 149)	0.95 ± 0.03 (0.84–0.99)	0.95 ± 0.06 (0.34–0.99)	0.97 ± 0.01 (0.92–0.99)	2.79 ± 3.07 (0.05–18.78)	7.43 ± 26.68 (0.08–326.36)	2.70 ± 2.35 (0.01–14.84)
	p value†	< 0.001	< 0.001	< 0.001	0.001	0.022	< 0.001
Disease type (n* = 426)	Gastric cancer (n = 95)	0.97 ± 0.02 (0.84–0.99)	0.98 ± 0.01 (0.94–0.99)	0.97 ± 0.01 (0.92–0.99)	2.32 ± 1.93 (0.02–10.51)	3.18 ± 2.28 (0.05–10.54)	1.94 ± 1.75 (0.02–10.11)
	Sepsis (n = 234)	0.95 ± 0.03 (0.75–0.99)	0.96 ± 0.05 (0.34–0.99)	0.97 ± 0.01 (0.93–0.99)	2.74 ± 3.08 (0.05–20.36)	5.06 ± 21.48 (0.04–326.36)	2.01 ± 1.92 (0.01–14.84)
	Healthy (n = 97)	0.97 ± 0.01 (0.91–0.99)	0.98 ± 0.02 (0.89–0.99)	0.97 ± 0.01 (0.90–0.99)	1.52 ± 1.48 (< 0.01–10.41)	3.55 ± 2.83 (0.03–18.25)	1.80 ± 1.66 (0.01–12.75)
	p value‡	< 0.001	< 0.001	0.644	< 0.001	0.318	0.521
	Contrasted pairings§	Cancer > sepsis (*p* < 0.001) Healthy > sepsis (*p* < 0.001)	Cancer > sepsis (*p* < 0.001) Healthy > sepsis (*p* < 0.001)	None	Sepsis > healthy (*p* < 0.001) Cancer > healthy (*p* = 0.003)	None	None

Data are presented as mean ± standard deviation, with range in parenthesis. *'n' indicates number of CT scans, †*p* values of independent *t* test, ‡*p* values of analysis of variance, §Pairings showing significant difference in Bonferroni corrected post-hoc *t* tests.

less than 10 cm^2 showing high CSA errors up to 326.36%. However, even human experts have difficulty in segmenting the target area accurately because the abdominal wall muscle and subcutaneous fat can be extremely thin and ill-demarcated in cachexic patients.

When subcutaneous edema is present, the X-ray attenuation of the subcutaneous fat area is increased due to high interstitial fluid content within the fatty tissue; this may cause difficulty in demarcating the boundary between subcutaneous fat and abdominal muscles (32, 33). Among the cases requiring manual adjustment of the FCN-based segmentation map, the majority (78.3%) showed diffuse subcutaneous edema. However, when subcutaneous fluid accumulation occurs in dependent regions such as the posterior back (i.e., dependent edema), even human experts may have difficulty in accurately segmenting the subcutaneous fat and muscle areas.

The conditions of muscle/fat wasting and subcutaneous edema, which may considerably influence the segmentation performance of FCN, are both closely related to disease type. In the internal validation dataset composed of gastric cancer patients, sepsis patients, and healthy subjects, the segmentation performance was both quantitatively and qualitatively the poorest in the sepsis group. The sepsis group showed the smallest muscle, subcutaneous fat, and visceral fat segmentation areas and had the highest proportion of patients with subcutaneous edema requiring manual adjustment (8.5%, 20 out of 234 CT scans). Sepsis may frequently accompany muscle wasting, such as in sepsis-induced myopathy, fat wasting, and subcutaneous edema, which possibly hindered accurate discrimination between muscle and fat.

Although FCN successfully generated results in all cases, some portions of the FCN-based segmentation were not qualitatively appropriate. Segmentation maps of 23 cases (3.9%) in the validation datasets showed errors in the segmentation of muscle or subcutaneous fat, mainly due to subcutaneous edema and presence of metallic artifacts from surgical devices. The majority of these errors occurred in the sepsis group (20 out of 234 CT scans); thus, our system should be further refined with regard to these difficult cases. We plan to train the deep learning system on larger datasets of more than 10000 patients from a broader range of clinical settings including sepsis, cancer patients

 https://doi.org/10.3348/kjr.2019.0470

undergoing chemotherapy, and liver cirrhosis patients awaiting transplantation.

As to the other observed causes of segmentation error, it is likely that metallic artifacts might cause under-segmentation by masking parts of muscle tissue (n = 3), whereas an enterostomy stump might cause over-segmentation by being incorrectly classified as muscle tissue (n = 1). In a case that required manual correction with marked intramuscular fat infiltration (n = 1), the intramuscular fat components were intermingled with muscle tissues and lowered the muscle density; thus, accurate discrimination between muscle and fat might have been limited. However, in all other cases with intramuscular fat infiltration, the segmentation was adequate and did not require manual correction. This issue is closely associated with the evaluation of muscle quality and myosteatosis, requiring further studies. In our current system, we could measure the muscle density in HUs from the segmented area; however, further studies are required with regard to muscle quality maps or myosteatosis indices.

Recently, several researchers showed acceptable performance in the automated segmentation of abdominal muscle and/or fat (18-22, 34, 35). These include Lee et al. (18), who showed reasonable performance using FCN-based automated segmentation of abdominal wall muscle and Weston et al. (21), with similarly good performance using U-Net based quantification of body composition. The performance of our FCN-based segmentation system for muscle (DSC: 0.96, CSA error: 2.26%) was a little higher than that of the Lee et al. (18) model (DSC: 0.93, CSA error: 3.68%). The larger sample size and more varied pathologies included in the training data may have improved the performance of our FCN-based segmentation system in comparison with theirs. Moreover, our FCN-based segmentation performance was equivalent or slightly higher than that of the Weston et al. (21) model, trained with 2430 CT examinations, in which the DSC values for muscle and visceral fat in their test dataset were 0.96 and 0.94, respectively. Before training the FCN-based segmentation system, we performed data augmentation generating 11167 training images, several data-adaptive preprocessing steps prior to assigning grayscale information to each pixel in consideration of the suboptimal consistency of HUs across different CT images, and we consider these preparation processes instrumental in achieving the high performance of our deep learning system.

There are some limitations to this study. First, this was a retrospective study, and the patient recruitment process was not consecutive; it depended on physicians referring their patients to the imaging core lab of our institution, and this may have resulted in selection bias. Second, although we included subjects with varied pathologies including cancer, sepsis, and healthy individuals, the study population may still not be representative of the general population. Furthermore, external validation was performed using data from a limited number of subjects from only two institutions; large-scale external validation might be necessary.

In conclusion, our FCN-based segmentation system exhibited high performance in the accurate segmentation of abdominal muscle and fat. Therefore, this fully automated segmentation system could be utilized in various clinical and research protocols for the quantitative analysis of body morphometry.

Supplementary Materials

The Data Supplement is available with this article at https://doi.org/10.3348/kjr.2019.0470.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

ORCID iDs

Kyung Won Kim
https://orcid.org/0000-0002-1532-5970
Hyo Jung Park
https://orcid.org/0000-0002-2364-9940
Yongbin Shin
https://orcid.org/0000-0003-2753-9586
Jisuk Park
https://orcid.org/0000-0003-0037-0274
Hyosang Kim
https://orcid.org/0000-0001-8140-9534
In Seob Lee
https://orcid.org/0000-0003-3099-0140
Dong-Woo Seo
https://orcid.org/0000-0001-8104-0247
Jimi Huh
https://orcid.org/0000-0002-8832-6165
Tae Young Lee
https://orcid.org/0000-0001-9983-2349

Korean Journal of Radiology
KJR

TaeYong Park
https://orcid.org/0000-0002-7523-8975
Jeongjin Lee
https://orcid.org/0000-0001-9676-271X

REFERENCES

1. Bosello O, Zamboni M. Visceral obesity and metabolic syndrome. *Obes Rev* 2000;1:47-56
2. Wajchenberg BL. Subcutaneous and visceral adipose tissue: their relation to the metabolic syndrome. *Endocr Rev* 2000;21:697-738
3. Blauwhoff-Buskermolen S, Versteeg KS, de van der Schueren MA, den Braver NR, Berkhof J, Langius JA, et al. Loss of muscle mass during chemotherapy is predictive for poor survival of patients with metastatic colorectal cancer. *J Clin Oncol* 2016;34:1339-1344
4. Kuroki LM, Mangano M, Allsworth JE, Menias CO, Massad LS, Powell MA, et al. Pre-operative assessment of muscle mass to predict surgical complications and prognosis in patients with endometrial cancer. *Ann Surg Oncol* 2015;22:972-979
5. Reisinger KW, van Vugt JL, Tegels JJ, Snijders C, Hulsewé KW, Hoofwijk AG, et al. Functional compromise reflected by sarcopenia, frailty, and nutritional depletion predicts adverse postoperative outcome after colorectal cancer surgery. *Ann Surg* 2015;261:345-352
6. Bokshan SL, Han AL, DePasse JM, Eltorai AE, Marcaccio SE, Palumbo MA, et al. Effect of sarcopenia on postoperative morbidity and mortality after thoracolumbar spine surgery. *Orthopedics* 2016;39:e1159-e1164
7. Jones K, Gordon-Weeks A, Coleman C, Silva M. Radiologically determined sarcopenia predicts morbidity and mortality following abdominal surgery: a systematic review and meta-analysis. *World J Surg* 2017;41:2266-2279
8. Fukuda Y, Yamamoto K, Hirao M, Nishikawa K, Nagatsuma Y, Nakayama T, et al. Sarcopenia is associated with severe postoperative complications in elderly gastric cancer patients undergoing gastrectomy. *Gastric Cancer* 2016;19:986-993
9. Boutin RD, Yao L, Canter RJ, Lenchik L. Sarcopenia: current concepts and imaging implications. *AJR Am J Roentgenol* 2015;205:W255-W266
10. Jones KI, Doleman B, Scott S, Lund JN, Williams JP. Simple psoas cross-sectional area measurement is a quick and easy method to assess sarcopenia and predicts major surgical complications. *Colorectal Dis* 2015;17:020-026
11. Mourtzakis M, Prado CM, Lieffers JR, Reiman T, McCargar LJ, Baracos VE. A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. *Appl Physiol Nutr Metab* 2008;33:997-1006
12. Shuster A, Patlas M, Pinthus JH, Mourtzakis M. The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. *Br J Radiol* 2012;85:1-10
13. McDonald AM, Swain TA, Mayhew DL, Cardan RA, Baker CB, Harris DM, et al. CT measures of bone mineral density and muscle mass can be used to predict noncancer death in men with prostate cancer. *Radiology* 2017;282:475-483
14. Shen W, Punyanitya M, Wang Z, Gallagher D, St-Onge MP, Albu J, et al. Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image. *J Appl Physiol (1985)* 2004;97:2333-2338
15. Kamiya N, Zhou X, Chen H, Muramatsu C, Hara T, Yokoyama R, et al. Automated segmentation of psoas major muscle in X-ray CT images by use of a shape model: preliminary study. *Radiol Phys Technol* 2012;5:5-14
16. Polan DF, Brady SL, Kaufman RA. Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study. *Phys Med Biol* 2016;61:6553-6569
17. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. *Radiographics* 2017;37:2113-2131
18. Lee H, Troschel FM, Tajmir S, Fuchs G, Mario J, Fintelmann FJ, et al. Pixel-level deep segmentation: artificial intelligence quantifies muscle on computed tomography for body morphometric analysis. *J Digit Imaging* 2017;30:487-498
19. Burns JE, Yao J, Chalhoub D, Chen JJ, Summers RM. A machine learning algorithm to estimate sarcopenia on abdominal CT. *Acad Radiol* 2019. pii: S1076-6332(19)30165-5
20. Decazes P, Tonnelet D, Vera P, Gardin I. Anthropometer3D: automatic multi-slice segmentation software for the measurement of anthropometric parameters from CT of PET/CT. *J Digit Imaging* 2019;32:241-250
21. Weston AD, Korfiatis P, Kline TL, Philbrick KA, Kostandy P, Sakinis T, et al. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology* 2019;290:669-679
22. Hu P, Huo Y, Kong D, Carr JJ, Abramson RG, Hartley KG, et al. Automated characterization of body composition and frailty with clinically acquired CT. *Comput Methods Clin Appl Musculoskelet Imaging (2017)* 2018;10734:25-35
23. Prado CM, Lieffers JR, McCargar LJ, Reiman T, Sawyer MB, Martin L, et al. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: a population-based study. *Lancet Oncol* 2008;9:629-635
24. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* 2017;39:640-651
25. Weisstein EW. Affine transformation. Wolfram MathWorld Web site. http://mathworld.wolfram.com/AffineTransformation.html. Published August 5, 2018. Accessed August 14, 2019
26. Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol* 2017;90:20160665
27. Perona P, Malik J. Scale-space and edge detection using

 https://doi.org/10.3348/kjr.2019.0470

anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell* 1990;12:629-639
28. Cropp RJ, Seslija P, Tso D, Thakur Y. Scanner and kVp dependence of measured CT numbers in the ACR CT phantom. *J Appl Clin Med Phys* 2013;14:4417
29. Vala HJ, Baxi A. A review on Otsu image segmentation algorithm. *IJARCET* 2013;2:387-389
30. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern Syst* 1979;9:62-66
31. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945;26:297-302
32. Callahan LA, Supinski GS. Sepsis-induced myopathy. *Crit Care Med* 2009;37:S354–S367
33. Hotchkiss RS, Moldawer LL, Opal SM, Reinhart K, Turnbull IR, Vincent JL. Sepsis and septic shock. *Nat Rev Dis Primers* 2016;2:16045
34. Bridge CP, Rosenthal M, Wright B, Kotecha G, Fintelmann F, Troschel F, et al. *Fully-automated analysis of body composition from CT in cancer patients using convolutional neural networks*. In: Stoyanov D, Taylor Z, Sarikaya D, McLeod J, Ballester MAG, Codella NCF, eds. *OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*. Cham: Springer, 2018:204-213
35. Popuri K, Cobzas D, Esfandiari N, Baracos V, Jägersand M. Body composition assessment in axial CT images using FEM-based automatic segmentation of skeletal muscle. *IEEE Trans Med Imaging* 2016;35:512-520

https://doi.org/10.3348/kjr.2019.0470