International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-021-02345-w

ORIGINAL ARTICLE

Automated segmentation of an intensity calibration phantom in clinical CT images using a convolutional neural network

Keisuke Uemura[1,2] · Yoshito Otake[1] · Masaki Takao[3] · Mazen Soufi[1] · Akihiro Kawasaki[1] · Nobuhiko Sugano[2] · Yoshinobu Sato[1]

Received: 21 December 2020 / Accepted: 4 March 2021

Abstract

Purpose In quantitative computed tomography (CT), manual selection of the intensity calibration phantom's region of interest is necessary for calculating density (mg/cm^3) from the radiodensity values (Hounsfield units: HU). However, as this manual process requires effort and time, the purposes of this study were to develop a system that applies a convolutional neural network (CNN) to automatically segment intensity calibration phantom regions in CT images and to test the system in a large cohort to evaluate its robustness.

Methods This cross-sectional, retrospective study included 1040 cases (520 each from two institutions) in which an intensity calibration phantom (B-MAS200, Kyoto Kagaku, Kyoto, Japan) was used. A training dataset was created by manually segmenting the phantom regions for 40 cases (20 cases for each institution). The CNN model's segmentation accuracy was assessed with the Dice coefficient, and the average symmetric surface distance was assessed through fourfold cross-validation. Further, absolute difference of HU was compared between manually and automatically segmented regions. The system was tested on the remaining 1000 cases. For each institution, linear regression was applied to calculate the correlation coefficients between HU and phantom density.

Results The source code and the model used for phantom segmentation can be accessed at https://github.com/keisuke-uemura/CT-Intensity-Calibration-Phantom-Segmentation. The median Dice coefficient was 0.977, and the median average symmetric surface distance was 0.116 mm. The median absolute difference of the segmented regions between manual and automated segmentation was 0.114 HU. For the test cases, the median correlation coefficients were 0.9998 and 0.999 for the two institutions, with a minimum value of 0.9863.

Conclusion The proposed CNN model successfully segmented the calibration phantom regions in CT images with excellent accuracy.

Keywords Artificial intelligence · Bone mineral density · Deep learning · Quantitative computed tomography · Phantom segmentation · U-net

Introduction

Quantification of bone mineral density (BMD) is necessary in the diagnosis of osteopenia and osteoporosis. Usually, lumbar vertebral or proximal femoral BMD is quantified using dual-energy X-ray absorptiometry—the procedure recommended by the World Health Organization and by several guidelines [1–3]. Yet, BMD assessment in other specific regions is also important, as it can be used for surgical planning to achieve good clinical results; to this end, studies have used quantitative computed tomography (CT) images to determine local BMD in the proximal femur [4], femoral head [5, 6], and distal radius [7].

✉ Keisuke Uemura
keisuke-uemura@is.naist.jp

[1] Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara, Japan

[2] Department of Orthopaedic Medical Engineering, Osaka University Graduate School of Medicine, Suita, Osaka, Japan

[3] Department of Orthopaedics, Osaka University Graduate School of Medicine, Suita, Osaka, Japan

Published online: 17 March 2021

To quantify BMD using CT, an intensity calibration phantom that contains known densities of either hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$, or dipotassium hydrogen phosphate, K_2HPO_4, must be included in the field of view (FOV) to be able to convert radiodensity (in Hounsfield units [HU]) to bone density (in mg/cm^3). This conversion is necessary when comparing CT results between patients and between institutions because studies have shown that the type of CT device, imaging protocol (e.g. tube voltage and slice thickness), and reconstruction protocol (e.g. convolutional kernel) affect HU values [8–10]. Conventionally, researchers manually select calibration phantom regions of interest on CT images, measure the radiodensity within each region of interest, and apply a linear regression model to convert the radiodensity values into tissue density values [4–6, 10]; however, this process requires effort and is time-consuming, which should be avoided in multicenter studies with large datasets.

In this study, we aimed (1) to develop a system that automatically segments the calibration phantom's regions of interest in CT images and converts HU into mg/cm^3, and (2) to evaluate the accuracy and robustness of the system by using CT images acquired at different institutions.

Materials and methods

A total of 1040 cases, data from patients who underwent hip surgery at two institutions ($n=520$ each, denoted herein as hospitals A and B) were included in this retrospective study. Ethical approval was obtained from the Institutional Review Board of each participating hospital. In hospital A, the primary reasons for hip surgery were osteoarthritis ($n=390$), osteonecrosis ($n=69$), and implant loosening ($n=26$), and in hospital B, the primary reason was proximal femoral fracture ($n=511$). At both hospitals, preoperative CT images are routinely acquired with an intensity calibration phantom (B-MAS200, Kyoto Kagaku, Kyoto, Japan) placed under the patient's body, approximately under the hip (Fig. 1a). This phantom is made of urethane foam (0 mg/cm^3) and contains four hydroxyapatite rods with known densities (50 mg/cm^3, 100 mg/cm^3, 150 mg/cm^3, and 200 mg/cm^3). The manufacturer and model of the CT device and the imaging protocols used in hospitals A and B are shown in Table 1. The CT image matrix and voxel size were similar between the hospitals. All consecutive cases in which preoperative CT images were acquired with the calibration phantom were analyzed, with no exclusion criteria (hospital A: July 2012–May 2019; hospital B: February 2015–September 2020). Quality control of the CT scanners at each hospital was performed every 6 months by the manufacturer.

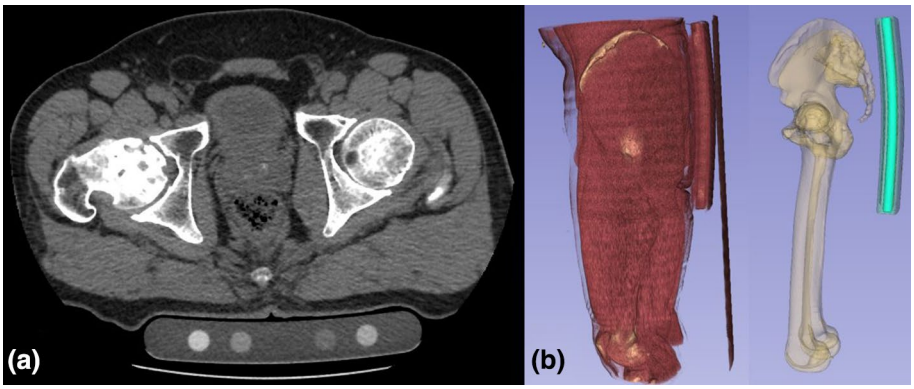

Fig. 1 An intensity calibration phantom, placed under the hip, that was included (**a**) in the field of view of an axial CT image at the level of the centre of the femoral head and (**b**) in a lateral view of volume rendering of the CT images (left) and segmented bones (right) shows deformation as a result of the patient's weight

Table 1 CT equipment, imaging protocols, and image characteristics of hospitals A and B

Hospital	CT manufacturer (model)	Tube voltage (kVp)	Convolution kernel (type)	Matrix size	Voxel size (mm)
A	General Electric (Optima CT660)	120	Standard (soft tissue)	512×512	(0.703–0.977) ×(0.703–0.977) ×(1.0–2.5)
B	Toshiba (Activion16)		FC30 (bone)		(0.622–0.972) ×(0.622–0.972) ×(0.5–2.0)

Segmentation of the calibration phantom

Because of the physical flexibility of the materials of which the calibration phantom is composed, the phantom deforms under the patient's weight (Fig. 1b). Thus, although the same model of calibration phantom was used throughout the study, segmentation of the CT image using simple rigid registration of the 3D phantom model was not applicable. Instead, we employed Bayesian U-Net [11], a convolutional neural network (CNN) for semantic segmentation. The training dataset consisted of 40 randomly selected cases (20 from each hospital), and in each case, the calibration phantom was manually segmented on all axial CT slices on which it appeared using Synapse Vincent software (v4.4, Fujifilm, Tokyo, Japan). Segmentation was performed by an orthopedic surgeon with 12 years of research experience (KU).

Image preprocessing and details of model training

As preprocessing, the intensity of the CT volumes was normalized to map [− 400, 700] HU to [0, 255] (i.e. intensities smaller than 400 HU and larger than 700 HU were represented as 0 and 255, respectively). During the training phase, data augmentation was performed so that the model would be invariant to the FOV of the scan and the inclination of the phantom. Specifically, translation by [− 10%, + 10%] of the matrix size, rotation of [− 15°, + 15°], scaling of [− 20%, + 20%], and shear transformation by a shear angle of $[-\pi/4 + \pi/4]$ radians, were performed. For the model training, the weights were initialized as was reported by He K et al. [12], and optimized using adaptive moment estimation (Adam [13]) for 1×10^5 iterations with an initial learning rate of 0.0001. The batch size was 2.

Automated calibration and postprocessing

Analysis of the CNN's segmentation results was performed on the remaining 1000 cases ($n = 500$ from each hospital). After the regions of the calibration phantom were defined on each axial slice (Fig. 2a), the segmented regions were eroded using a 3-pixel disk-shaped structuring element to avoid susceptibility to small variations at the boundaries that can affect the radiodensity values measured in each material

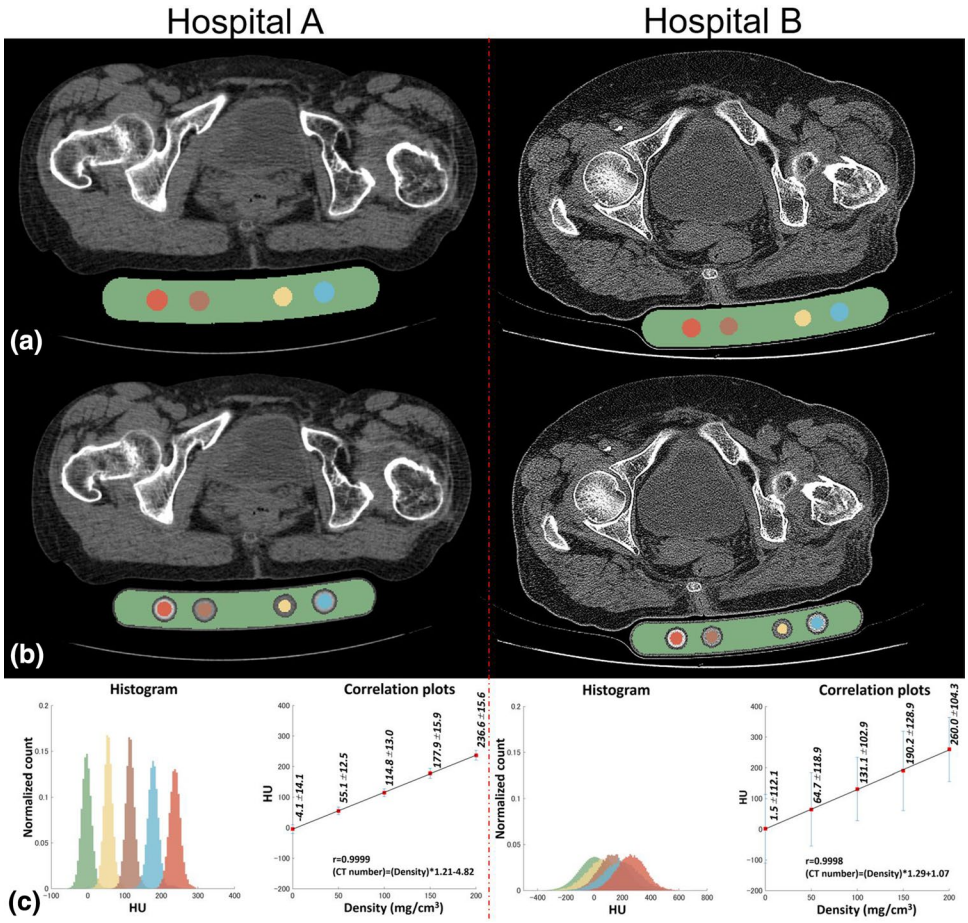

Fig. 2 Example CT images from hospital A (left) and hospital B (right) show (**a**) the intensity calibration phantom segmented into regions representing 0 mg/cm^3, 50 mg/cm^3, 100 mg/cm^3, 150 mg/cm^3, and 200 mg/cm^3, indicated by green, yellow, brown, cyan, and vermilion, respectively, using Bayesian U-Net, and (**b**) the regions in the image filled and eroded. (**c**) Example histograms of each region's radiodensity (left), and linear regression model, correlation coefficients, and equation (right)

(Fig. 2b). Linear regression (the standard protocol suggested by the manufacturer) was applied to model the relationships between signal intensity [HU] and density [mg/cm^3] for each test case (Fig. 2c), and then the slopes and correlation coefficients of the regression models were calculated. MATLAB (v9.8, The MathWorks, Natick, MA, USA) was used for the postprocessing and calibration processes.

Quantitative assessment of segmentation accuracy

To assess segmentation accuracy, fourfold cross-validation was performed on the training dataset: 15 cases from each hospital ($n = 30$ total) were randomly selected for training, and the remaining 5 cases from each hospital ($n = 10$ total) were used for validation in each fold. Accuracy was evaluated using the Dice coefficient [14] and average symmetric surface distance (ASD) [15]. Furthermore, to determine the effects of the differences in the segmentation method, the absolute differences of radiodensity values were compared between the manually and automatically segmented regions.

Calculation of density using the regression models

To analyze the effects of calibration, the linear regression models were applied in the range between − 100 and 700 HU (the radiodensity range of human tissues often used as the target of clinical analysis), and the results were compared between the two hospitals at every integer HU value (i.e. a total of 801 statistical comparisons).

Segmentation in cases including metal implants and cases with the phantom partially located outside of the FOV

As a sub-analysis, the segmentation accuracy in cases that may degrade the performance of the developed system was evaluated. Specifically, we compared the cases with metal implants in the FOV ($n = 273$, group B), and cases with the phantom located partially outside of the FOV (groups C and D) against all other cases (group A). Cases that had the phantom outside of the FOV were grouped according to the percentage of volume outside the FOV: > 20% (group C, $n = 44$) and > 10% (group D, $n = 119$).

Comparison between the conventional manual method and the automated method

The regression models of the automated method were compared with the results of the conventional manual method in 100 randomly selected cases (50 from each hospital). First, circular-shaped regions of interest were manually defined for each rod on three axial CT slices by one researcher (KU). These three slices were randomly selected from CT images that did not have obvious halation on each rod. Then, a linear regression model was generated using the mean values of the three slices (Fig. 3), and the correlation coefficients were compared with those of the automated method. To assess inter- and intra-observer reliability, KU performed the manual procedure twice, and another orthopedic surgeon (MT, with 22 years of research experience) also performed the manual procedure. The intraclass correlation coefficient (ICC) was quantified by comparing the HU values measured in each selected region.

Statistical analysis

Normality was assessed with a Shapiro–Wilk test. Data were expressed as mean ± standard deviation when normally distributed and as median (interquartile range) when not normally distributed. Data were compared using the Mann–Whitney U-test when not normally distributed. The Benjamini–Hochberg procedure was used to correct for multiple comparisons. The Wilcoxon signed-rank test was used to compare paired, non-normally distributed data. All statistical analyses were performed using MATLAB, and values of $p < 0.05$ were considered statistically significant.

Results

Segmentation accuracy

Automated segmentation was successful in all 1000 test cases. The source code and the model used for segmenting the phantom are open source and can be accessed via https://github.com/keisuke-uemura/CT-Intensity-Calibration-Phantom-Segmentation.

After fourfold cross-validation, the median Dice coefficient was 0.977 (0.023), and the median ASD was 0.116 mm (0.108 mm) (Fig. 4). No significant difference was found between the hospitals' Dice coefficients ($p = 0.84$), but the ASD of hospital A was significantly larger than that of hospital B ($p = 0.02$) (Table 2). When the HU in the segmented regions were compared between the manual method and the automated method, difference of 0.142 HU (0.280) was found for hospital A and difference of 0.082 HU (0.124) was found for hospital B. The median absolute difference for all cases was 0.114 HU (0.158).

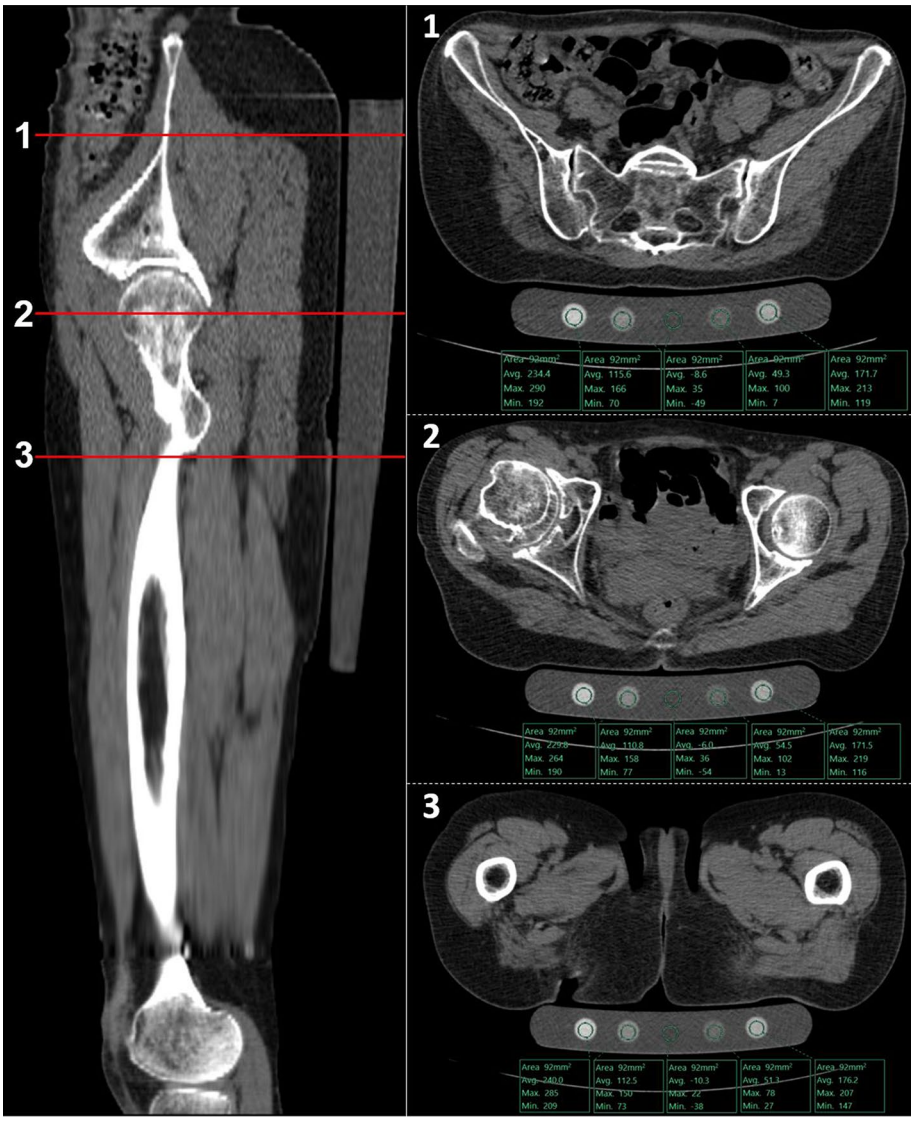

Fig. 3 Manual measurement of calibration phantom radiodensity was performed on three axial slices (right) indicated by red horizontal lines labelled 1, 2, and 3 on the sagittal view (left)

Comparison between regression models of each hospital

The median correlation coefficients of the regression were 0.9998 (0.0004) and 0.9999 (0.0001) for hospitals A and B, respectively. The median slope of the regression model was 0.841 (0.027) and 0.744 (0.041) for hospitals A and B, respectively (Fig. 5), which was significantly different ($p < 0.001$). Within the range between − 100 and 700 HU, there were significant differences between the hospitals for the ranges between − 100 and 4 HU and between 11 and 700 HU: the difference in the result for tissue density was 0.6 mg/cm^3 at 0 HU and 58.2 mg/cm^3 at 600 HU (Fig. 5).

Segmentation in cases with metal implants and cases with the phantom partially located outside of the FOV

Automated segmentation was possible in cases with severe artefacts resulting from metal implants in the FOV and cases in which the phantom was located partially outside the FOV (Fig. 6). The median correlation coefficient for both groups was 0.9999 (Table 3), but groups B and D had significantly smaller coefficients than group A ($p < 0.001$ and $p = 0.03$, respectively) (Table 3).

Comparison between the conventional manual method and the automated method

For the 100 cases that were compared, the median correlation coefficient of the regression models was 0.9996 (0.0010) and 0.9999 (0.0003) for the conventional manual and

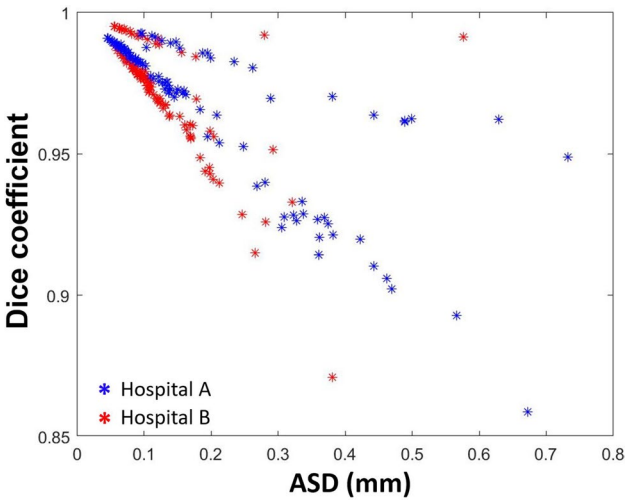

Fig. 4 Scatter plots of the Dice coefficient and average symmetric surface distance (ASD) for all five regions of the calibration phantom. Results of hospitals A and B are indicated in blue and red, respectively

automated methods, respectively. The automated method's correlation coefficient was significantly higher ($p < 0.01$). The intra- and inter-observer reliability (evaluated by ICC) were both 0.998, indicating excellent reliability.

Discussion

We applied a CNN to automatically segment the corresponding regions of differing radiodensity that corresponded to the different known tissue densities of an intensity calibration phantom used in clinical CT images. The model's accuracy, indicated by the Dice coefficient (0.977) and ASD (0.116 mm), and its robustness, indicated by the overall absolute difference between manual and automated segmentation (0.114 HU), were excellent after training. However, significant differences in calculated tissue density were found between the two hospitals, especially at larger values in the typical clinical range.

In quantitative CT analyses, researchers typically select the calibration phantom manually from a few axial slices to create a regression model between radiodensity and tissue density. Both the conventional manual and automated methods' correlation coefficients exceeded 0.999. As the correlation coefficient of the automated method was significantly higher, the automated method seems to be at least equivalent to the conventional manual method for developing calibration regression models. Because the automated method does not require a manual process (i.e. selection of the target axial slices and regions of interest), the automated system saves time and effort. It is therefore more suitable for multicenter studies in which many institutions and researchers participate.

There were significant differences between the tissue densities calculated with the regression models of hospitals A and B when CT image radiodensity ranged from − 100 to 4 and from 11 to 700. Small differences, even if statistically significant, may not be clinically important. However, because of the regression models' difference in slope (Fig. 5), the differences in calculated tissue density increased as the radiodensity value increased, resulting in a difference of 58.2 mg/cm^3 at 600 HU (Fig. 5). This finding is in line

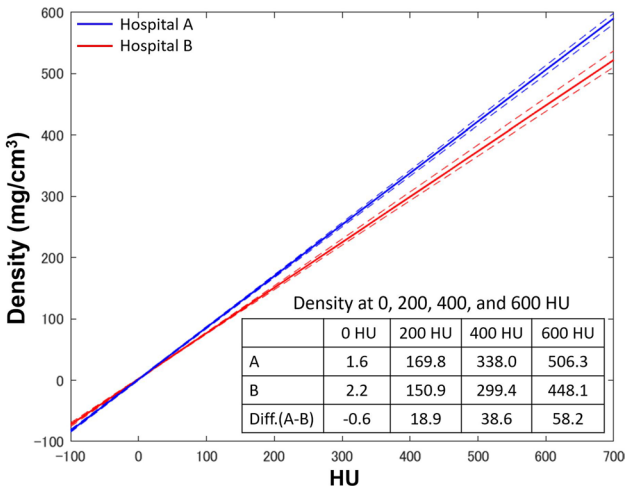

Fig. 5 Relationship between radiodensity (horizontal axis) and tissue density (vertical axis) in the models for hospitals A (blue) and B (red) from − 100 to 700 HU. Solid lines indicate the medians, and dotted lines indicate the interquartile ranges. The table in the lower right corner shows tissue densities calculated for 0 HU, 200 HU, 400 HU, and 600 HU with each hospital's model. Diff.: Difference

Table 2 Results of fourfold cross-validation

Parameter	Overall	Hospital A	Hospital B	p value
Dice coefficient	0.977 (0.023)	0.977 (0.025)	0.977 (0.017)	0.84
ASD (mm)	0.116 (0.108)	0.136 (0.208)	0.106 (0.073)	0.02
Absolute difference in HU	0.114 (0.158)	0.142 (0.280)	0.082 (0.124)	0.003

Data are expressed as median (interquartile range)

ASD average symmetric surface distance, *HU* Hounsfield units

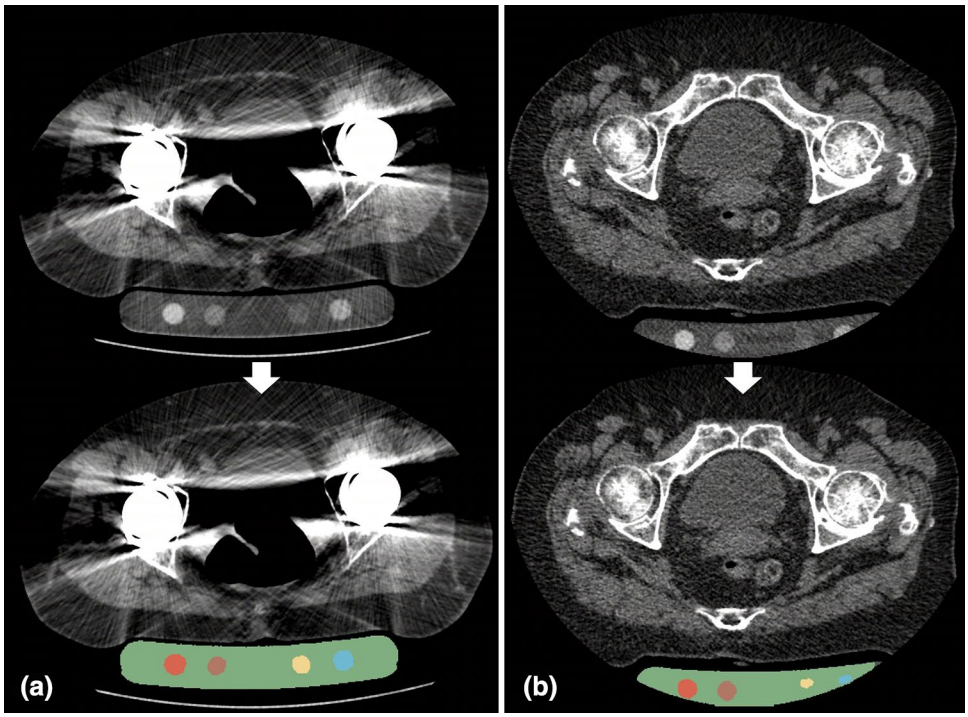

Fig. 6 Two examples of challenging cases: (**a**) image artifacts because of bilateral metallic hip implants and (**b**) an image in which the calibration phantom is located partially outside of the FOV

Table 3 Results of cases with metal implants and/or with the phantom located partially outside of the field of view (FOV)

Group	Number of cases (hospital A, hospital B)	Correlation coefficient	*p* value
A	636 (A: 295, B: 341)	0.9999 (0.0002)	N/A
B	273 (A: 186, B: 87)	0.9999 (0.0003)	< 0.001*
C	44 (A: 8, B: 36)	0.9999 (0.0001)	0.20
D	119 (A: 31, B: 88)	0.9999 (0.0002)	0.03*

Group A: No metal implants and no phantom regions located outside the FOV, Group B: With metal implants, Group C: Phantom located outside the FOV (> 20%), Group D: Phantom located outside the FOV (> 10%)

Data are expressed as median (interquartile range)

Total size of each group does not equal 1000, as some cases had both metal implants and the phantom located outside the FOV

*Indicates significant difference (significantly smaller than group A)

with the results of Giambini et al. [9], who found that large errors are expected if the static range definition of 300–600 HU is used to define cortical bone in CT images, as has been done in previous bone surface modelling and finite element studies [16–19]. We suggest that static definitions should not be used between institutes and recommend using an intensity calibration phantom when comparing BMD values between institutes.

Recently, studies have employed CNNs to diagnose hip diseases [20, 21] and to segment musculoskeletal regions in CT images [11]. To the best of our knowledge, no study has developed an automated system that uses a CNN to segment the regions of intensity calibration phantoms in CT images. Although the CNN used in this study has been reported previously for segmentation of musculoskeletal regions, we believe that the method's applicability to other target objects and in clinical workflows had not been evident, further emphasizing this study's contribution and importance. In recent studies, a phantom-less calibration method using internal reference tissues of each patient, such as the aortic blood tissue, pelvic visceral adipose tissue, muscle, and fat, has been reported [10, 22], with conflicting results. One paper reported the usefulness of the method [10], but the other recommended caution [22]. Importantly, these previous studies included only a limited number of cases because manual selection of the reference and regions of interest was necessary. It would be interesting to apply the system developed in this study to clarify the usefulness/accuracy of the phantom-less calibration method.

Manual effort is still necessary to quantify BMD from CT images because the bone regions of interest must be selected manually. As this process is time-consuming and error-prone [23], in future studies, we aim to develop a CNN to isolate the region of interest (e.g., femoral neck and spine) and thereby create a fully automated system, which would pave the way for multicenter quantitative CT studies with large datasets.

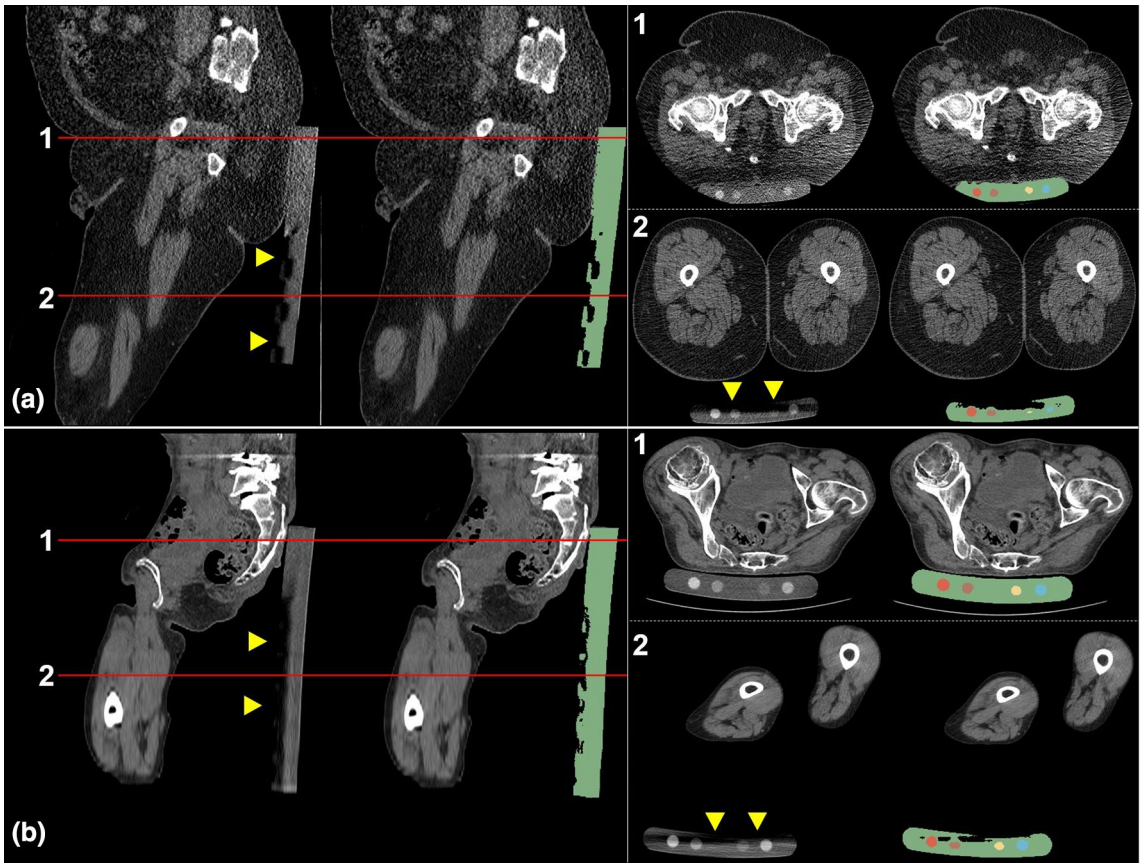

Fig. 7 CT images and segmentation results of the calibration phantom (sagittal and axial views) in two cases that with correlation coefficients: the models with the (**a**) lowest correlation coefficient (0.9863) and (**b**) second lowest correlation coefficient (0.9908) among the test cases. In (**a**), the border of the calibration phantom was difficult to distinguish on the axial slice at the femoral head level (right upper row, red horizontal line labelled 1 on the sagittal view) because of obesity (body mass index: 40.6), and halation (yellow triangles) on the axial slice at the mid-femur (red horizontal line labelled 2 on the sagittal view) caused segmentation errors. In (**b**), halation at the mid-femur caused segmentation deficiency

Limitations

This study had some limitations. First, although the system was tested at two hospitals with different CT devices, imaging protocols, and reconstruction protocols, including two types of convolutional kernels (i.e. soft tissue and bone) that are commonly used in the field of orthopedics, results may vary if CT images are acquired in different situations. However, it is likely that the system would be able to perform sufficiently if an appropriate training dataset were added. Second, because only axial slices were used, images with halation (e.g., metal artefacts and beam hardening) were included (Fig. 7). However, the effect of halation on the regression models was weakened (because there were relatively few of these images) and was likely negligible. This assumption is supported by the high minimum correlation coefficient (0.9863) and the median correlation coefficient of 0.9999 in the sub-analysis for each group (Table 3). Finally, as the training data were created based on one researcher's segmentation, the results may have varied if a different researcher created the training data. However, as the inter-observer reliability in measuring the radiodensity from the phantom was extremely high, it is very likely that only small differences would be found in such circumstances.

Conclusions

The CNN was able to accurately segment the intensity calibration phantom from the CT images, with a mean Dice coefficient of 0.977, ASD of 0.116 mm, and mean error of 0.114 HU. The median correlation coefficient of the regression models was > 0.999, which demonstrates the developed system's excellent ability to convert radiodensity into tissue density. Large differences in tissue density were found between the models in the range defined for cortical bone

(300–600 HU), indicating the necessity of using a calibration phantom to compare the results between institutions.

Acknowledgements This study was supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (KAKENHI) numbers 19H01176 and 20H04550. The authors thank Tatsuya Kitaura MD and the radiological technologists for their help with data acquisition.

Author contributions KU and YO contributed to conceptualization and methodology; KU, YO, and MS were involved in code writing; KU and AK contributed to formal analysis and investigation; KU contributed to writing—original draft preparation; YO, MT, MS, NS, and YS contributed to writing—review and editing; YO and YS contributed to funding acquisition. All authors read and approved the final manuscript.

Funding This study was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) Numbers 19H01176 and 20H04550.

Data availability The model used for phantom segmentation can be accessed via https://github.com/keisuke-uemura/CT-Intensity-Calibration-Phantom-Segmentation

Code availability The code used for phantom segmentation can be accessed via https://github.com/keisuke-uemura/CT-Intensity-Calibration-Phantom-Segmentation

Declarations

Conflict of interest The authors have nothing to disclose.

Ethics approval All procedures performed in this study were performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Consent to participate This study was approved by the Institutional Review Board of each participating hospital, and written informed consent was waived because of the retrospective design.

References

1. Kanis JA, Cooper C, Rizzoli R, Reginster JY (2019) European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int 30(1):3–44. https://doi.org/10.1007/s00198-018-4704-5
2. Orimo H, Nakamura T, Hosoi T, Iki M, Uenishi K, Endo N, Ohta H, Shiraki M, Sugimoto T, Suzuki T, Soen S, Nishizawa Y, Hagino H, Fukunaga M, Fujiwara S (2012) Japanese 2011 guidelines for prevention and treatment of osteoporosis–executive summary. Arch Osteoporos 7(1–2):3–20. https://doi.org/10.1007/s11657-012-0109-9
3. Camacho PM, Petak SM, Binkley N, Diab DL, Eldeiry LS, Farooki A, Harris ST, Hurley DL, Kelly J, Lewiecki EM, Pessah-Pollack R, McClung M, Wimalawansa SJ, Watts NB (2020) American Association of clinical endocrinologists/American college of endocrinology clinical practice guidelines for the diagnosis and treatment of postmenopausal osteoporosis-2020 update. Endocr Pract Off J Am Coll Endocr Am Assoc Clin Endocr 26(Suppl 1):1–46. https://doi.org/10.4158/gl-2020-0524suppl
4. Maeda Y, Sugano N, Saito M, Yonenobu K (2011) Comparison of femoral morphology and bone mineral density between femoral neck fractures and trochanteric fractures. Clin Orthop Relat Res 469(3):884–889. https://doi.org/10.1007/s11999-010-1529-8
5. Uemura K, Takao M, Otake Y, Hamada H, Sakai T, Sato Y, Sugano N (2018) The distribution of bone mineral density in the femoral heads of unstable intertrochanteric fractures. J Orthop Surg 26(2):2309499018778325. https://doi.org/10.1177/2309499018778325
6. Whitmarsh T, Otake Y, Uemura K, Takao M, Sugano N, Sato Y (2019) A cross-sectional study on the age-related cortical and trabecular bone changes at the femoral head in elderly female hip fracture patients. Sci Rep 9(1):305. https://doi.org/10.1038/s41598-018-36299-y
7. Hanusch BC, Tuck SP, Mekkayil B, Shawgi M, McNally RJQ, Walker J, Francis RM, Datta HK (2020) Quantitative computed tomography (QCT) of the distal forearm in men using a spiral whole-body CT scanner: description of a method and reliability assessment of the QCT Pro software. J Clin Densitom Off J Int Soc Clin Densitom 23(3):418–425. https://doi.org/10.1016/j.jocd.2019.05.005
8. Adams JE (2009) Quantitative computed tomography. Eur J Radiol 71(3):415–424. https://doi.org/10.1016/j.ejrad.2009.04.074
9. Giambini H, Dragomir-Daescu D, Huddleston PM, Camp JJ, An KN, Nassr A (2015) The effect of quantitative computed tomography acquisition protocols on bone mineral density estimation. J Biomech Eng 137(11):114502. https://doi.org/10.1115/1.4031572
10. Lee DC, Hoffmann PF, Kopperdahl DL, Keaveny TM (2017) Phantomless calibration of CT scans for measurement of BMD and bone strength-Inter-operator reanalysis precision. Bone 103:325–333. https://doi.org/10.1016/j.bone.2017.07.029
11. Hiasa Y, Otake Y, Takao M, Ogawa T, Sugano N, Sato Y (2020) Automated muscle segmentation from clinical CT using Bayesian U-net for personalized musculoskeletal modeling. IEEE Trans Med Imaging 39(4):1030–1040. https://doi.org/10.1109/tmi.2019.2940555
12. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. arXiv:150201852
13. Kingma DP, J B (2017) Adam: a method for stochastic optimization. arXiv:14126980
14. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302. https://doi.org/10.2307/1932409
15. Styner M, Lee J, Chin B, Chin M, Commowick O, Tran H, Markovic-Plese S, Jewells V, Warfield S (2008) 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. Midas J 1–5
16. Aamodt A, Kvistad KA, Andersen E, Lund-Larsen J, Eine J, Benum P, Husby OS (1999) Determination of Hounsfield value for CT-based design of custom femoral stems. J Bone Joint surg Br 81(1):143–147
17. Gausden EB, Nwachukwu BU, Schreiber JJ, Lorich DG, Lane JM (2017) Opportunistic use of CT imaging for osteoporosis screening and bone density assessment: a qualitative systematic review. J Bone Joint Surg Am 99(18):1580–1590. https://doi.org/10.2106/jbjs.16.00749
18. Kitamura K, Fujii M, Utsunomiya T, Iwamoto M, Ikemura S, Hamai S, Motomura G, Todo M, Nakashima Y (2020) Effect of sagittal pelvic tilt on joint stress distribution in hip dysplasia: a finite element analysis. Clin Biomech 74:34–41. https://doi.org/10.1016/j.clinbiomech.2020.02.011
19. Schreiber JJ, Anderson PA, Rosas HG, Buchholz AL, Au AG (2011) Hounsfield units for assessing bone mineral density and strength: a tool for osteoporosis management. J Bone Joint Surg Am 93(11):1057–1063. https://doi.org/10.2106/jbjs.j.00160

20. Mawatari T, Hayashida Y, Katsuragawa S, Yoshimatsu Y, Hamamura T, Anai K, Ueno M, Yamaga S, Ueda I, Terasawa T, Fujisaki A, Chihara C, Miyagi T, Aoki T, Korogi Y (2020) The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. Eur J Radiol 130:109188. https://doi.org/10.1016/j.ejrad.2020.109188
21. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, Chung IF, Liao CH (2019) Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 29(10):5469–5477. https://doi.org/10.1007/s00330-019-06167-y
22. Therkildsen J, Thygesen J, Winther S, Svensson M, Hauge EM, Böttcher M, Ivarsen P, Jørgensen HS (2018) Vertebral bone mineral density measured by quantitative computed tomography with and without a calibration phantom: a comparison between 2 different software solutions. J Clin Densitom Off J Int Soc Clin Densitom 21(3):367–374. https://doi.org/10.1016/j.jocd.2017.12.003
23. Feit A, Levin N, McNamara EA, Sinha P, Whittaker LG, Malabanan AO, Rosen HN (2019) Effect of positioning of the region of interest on bone density of the hip. J Clin Densitom Off J Int Soc Clin Densitom. https://doi.org/10.1016/j.jocd.2019.04.002