学 位 論 文

Deep Learning Classification and Grad-CAM-based Visualization for Osteoporotic Lumbar Vertebral Fractures on Radiographs
（単純 X 線写真における骨粗鬆性腰椎椎体骨折の
深層学習分類と Grad-CAM に基づく分類根拠に関する検討）

小野　陽平

北海道大学大学院保健科学院
保健科学専攻保健科学コース

２０２３年度

Contents

Abstract

Introduction: Early diagnosis and initiation of treatment for fresh osteoporotic lumbar vertebral fractures (OLVF) are crucial to maintain the fresh OLVF patient's activities of daily living and quality of life. Magnetic resonance imaging (MRI) is generally performed to differentiate fresh and old OLVF. However, MRI is a high-cost exam that burdens patients with severe back pain by forcing them to maintain their body position during long examinations. Furthermore, it could be difficult to perform in an emergency. MRI should therefore be performed in appropriately selected patients with a high suspicion of fresh fractures. As radiography is the first-choice imaging examination for the diagnosis of OLVF, improving screening accuracy with radiographs will optimize the decision of whether MRI is performed. In recent years, deep learning methods such as convolutional neural network (CNN) and Gradient-weighted Class Activation Mapping (Grad-CAM) have been used to solve various problems in the field of medical imaging. One of the most important features of CNN is high image classification performance based on high feature extraction capability. Grad-CAM can visualize the basis of CNN classification, which may be able to deepen our understanding of the CNN classification process. This study aimed to evaluate a method to automatically determine the presence of OLVF and classify old and fresh OLVF using a CNN model with radiographs, and the areas of interest to our CNN model based on Grad-CAM.

Materials and Methods: 523 in Institution 1 and 140 in Institution 2 patients with suspected OLVF who underwent both lumbar vertebrae (LV) radiography and MRI were included. A total of 3481 LV images in Institution 1 for training, validation, and testing and 662 LV images in Institution 2 for external validation images were collected. Visual evaluation with MRI images by two radiologists determined the ground truth of LV conditions such as normal, old, and fresh OLVF. Automatic object detection with you only look once at version 5 (YOLOv5)

was trained to recognize each lumber vertebral body and used to create the sample images for CNN classification. Three CNNs, Resnet-50, DenseNet-161, and ResNeXt-50, were ensembled to determine the final classification result. The classification performance on the LV conditions was calculated. Grad-CAM images were quantitatively evaluated and analyzed for the areas of interest in image classification by CNN.

Results: The interobserver agreement value for visual evaluation by two radiologists was 0.801. The intraobserver agreement values for raters 1 and 2 were 0.821 and 0.861, respectively. The detection performance of YOLOv5 was mAP (0.5) of 0.995 and mAP (0.5: 0.95) of 0.993 for the validation dataset and mAP (0.5) of 0.982 and mAP (0.5: 0.95) of 0.835 for the test dataset. The accuracy, sensitivity, specificity, and area under the curve in receiver operating characteristic analysis were 0.894, 0.836, 0.920, and 0.912 in the test and 0.867, 0.674, 0.866, and 0.855 in the external validation, respectively. Grad-CAM images had higher pixel values around the center of the image. The CNN classification was correctly based on the characteristics of the vertebral body rather than on background areas. There was a definite difference in the areas of interest to our CNN model in each group, normal vertebra, old, and fresh OLVF.

Conclusions: The proposed CNN-based method demonstrated high performance in determining the presence of OLVF and classifying old or fresh OLVF on radiography without manual procedures. Utilizing objective classification results from our CNN is expected to improve the accuracy of fresh OLVF screening. This may lead to appropriate decisions on the indication for close examination with MRI. The quantitative evaluation of Grad-CAM images allowed us to identify the areas of interest for the CNN model created in this study, which were found to be mainly the anterior vertebral wall and endplates. Further detailed Grad-CAM analysis might provide new knowledge for OLVF evaluation with the human eye in clinical practice in the future.

Abbreviations

OLVF	osteoporotic lumbar vertebral fracture
ADL	activities of daily living
QOL	quality of life
MRI	magnetic resonance imaging
DL	deep learning
CNN	convolutional neural network
Grad-CAM	Gradient-weighted Class Activation Mapping
YOLOv5	you only look once at version 5
STARD	Standards for Reporting Diagnostic Accuracy Studies
FPD	flat panel detector
SID	source-to-image receptor distance
AEC	auto exposure control
NNC	Neural Network Console
mAP	mean average precision
IoU	intersection over union
ROC	receiver operating characteristic
CI	confidence intervals
SVM	support vector machine

Preface

This article describes a new study that attempted to determine the presence of osteoporotic lumbar vertebral fracture (OLVF) and classify old or fresh OLVF from radiographs by utilizing deep learning techniques. In this preface, we briefly describe the background and the significance of the study, and the deep learning techniques used.

In recent years, Japan's population has been aging rapidly, and the increase in the number of osteoporotic patients has become a significant problem. Osteoporosis is a very common disease that affects 20-30% of men and 30-50% of women at least once in their lifetime. One of the most common complications of osteoporosis is osteoporotic vertebral fracture, which is a serious problem in the health care of vertebral fracture patients, especially the elderly, because its progression significantly reduces the activities of daily living (ADL) and quality of life (QOL) of patients. Currently, the definitive diagnosis of new-onset vertebral fractures is mainly made by magnetic resonance imaging (MRI), but this is problematic due to resource and time constraints in clinical practice and the high burden on the patient. Considering the medical environment in Japan and the actual situation of vertebral fracture patients, we attempted to develop a new vertebral fractures evaluation method that combines deep learning techniques with radiographs, which are easy to obtain images with low burden. This method is expected to enable quicker and more efficient diagnosis of lumbar vertebral fractures by accurately identifying the fractured vertebra and its old or fresh without relying on experience or knowledge of physicians, and by providing objective indicators to physicians. In addition, as the number of osteoporosis patients is expected to continue to increase with further lengthening of life expectancy, the significance of this study will be further enhanced, considering that

early detection of vertebral fractures has a significant impact on patients' QOL and ADL.

Three deep learning techniques are used in this study. The first is automatic lumbar vertebrae detection and cropping. These deep learning techniques are very important in the creation of the input images to convolutional neural network (CNN) to shorten the time and eliminate manual procedures bias. In this process, each of the five lumbar vertebrae was recognized individually, and each vertebral body image was created from the original radiograph based on the detected coordinate information. The second is the classification of vertebral fracture conditions using CNN. This is the main task of this study. The format of this study was to classify the input vertebral body images into three classes: normal, old, and fresh fractures. In addition, a method was employed to determine the final result based on the output results of the three CNNs to further improve the classification accuracy. The third is Gradient-weighted Class Activation Mapping (Grad-CAM) image analysis. Generally, the problem with CNN-based image classification is that the basis for the classification is unclear. We believe that the evaluation of the important areas for classification will deepen our understanding of the CNN classification process, improve its reliability, and may provide new knowledge for diagnosis with the human eye in clinical practice.

The background and the deep learning techniques used in this study described in this chapter will be discussed in detail in subsequent chapters.

1. Introduction

Osteoporotic lumbar vertebral fracture (OLVF) is one of the most common complications in osteoporotic patients. The main factors associated with OLVF are a decrease in bone density, bone quality deterioration, and bone microstructure degeneration caused by osteoporosis [1,2]. Since the risk of osteoporosis increases with age, the number of patients with osteoporosis and those who develop OLVF will continue to increase as life expectancy increases [3–5].

As OLVF progresses, chronic severe pain, decreased vertebral body height, a round back, gait difficulty, decreased pulmonary function, and increased mortality significantly lead to a decrease in activities of daily living (ADL) and quality of life (QOL) [6,7]. Although the progression of these symptoms is often stabilized in patients with old OLVF, the prognosis of fresh OLVF is poor unless appropriate intervention is provided early. Therefore, when OLVF is confirmed, it is crucial to confirm the diagnosis of fresh OLVF as early as possible, relieve pain, and prevent the progression of crushing to maintain ADL and QOL in fresh OLVF patients [8–10].

Generally, magnetic resonance imaging (MRI) is used to determine whether OLVF is old or fresh [11]. The presence of fresh OLVF is indicated by low signal intensity on T1-weighted images and high signal intensity on T2-weighted and short tau inversion recovery images, which reflect vertebral edema. MRI is a highly sensitive and specific method to determine whether an OLVF is old or fresh [12]. MRI can also detect subclinical fractures and identify the site of injury. MRI is an important tool for diagnosing OLVF because it allows for more detailed treatment strategy decisions [13,14]. On the other hand, MRI is a high-cost exam that burdens patients with severe back pain

by forcing them to maintain their body position during long examinations. MRI also has drawbacks, such as the difficulty of emergency examinations and the limited number of examinations that can be performed in a day [11,13,15,16]. MRI should therefore only be performed on appropriately selected patients with a high suspicion of fresh OLVF who are most likely to require therapeutic intervention [15].

Currently, when fresh OLVF is suspected, the first-choice imaging examination is radiography, which is superior in image acquisition time, simplicity, low exposure dose, and low cost [10,17,18]. Radiography can screen for the presence of OLVF by capturing morphological changes in the vertebral body. However, due to its characteristics, it is difficult to diagnose fine fractures with few morphological changes immediately after onset and to accurately determine the stage of OLVF even when a decrease in vertebral height is observed [13,15,19,20]. It is also known that the probability of developing multiple OLVFs is higher in patients who have developed OLVF once in the past [21,22]. In patients with multiple OLVFs, especially when there are both old and fresh fractures in a radiograph, it is more difficult to visually identify the causative vertebrae from their morphology and determine whether MRI is indicated. In such cases, an old or fresh OLVF diagnosis may significantly depend on the physician's experience and ability. To optimize the decision of the indication for MRI, it is therefore necessary to improve the accuracy of OLVF screening in radiography and develop a new automatic evaluation method that is not easily influenced by differences in the physicians' experience and ability.

In recent years, deep learning (DL) methods based on network structures called convolutional neural networks (CNNs) have been used to solve various problems in the field of medical imaging. One of the most important features is high image classification performance based on high feature extraction capability [23,24]. In this study, we utilize

DL techniques of object detection, 3-class classification based on CNNs, and Gradient-weighted Class Activation Mapping (Grad-CAM) image analysis. The object detection algorithm used in this study is "you only look once" at version 5 (YOLOv5). It has been reported that YOLO has high efficiency and detection accuracy and meets the requirements for use in clinical practice [25–28]. Grad-CAM is a CNN-based technique to visualize the areas of interest in image classification by CNN [29]. Generally, the problem with CNN-based image classification is that the basis for the classification is unclear. Grad-CAM is the method that can solve that problem.

In a previous study that attempted to identify fresh vertebral compression fractures using radiography, the CNN model was designed for a 2-class classification of fresh and old fractures but could not identify normal vertebrae [30]. Therefore, it was necessary to manually select vertebrae with suspected fractures before applying the CNN model. To the best of our knowledge, this is the first study to be able to classify not only fresh and old vertebrae but also normal vertebrae. A 3-class classification allows all vertebrae to be included without prior selection. This reduces the burden of selecting the target vertebrae and the risk of missing a fresh fracture in another vertebra other than the target vertebra.

We have developed an efficient evaluation method with high detection capability by combining radiography's quick image acquisition with CNNs high image classification performance. The use of CNNs to accurately detect fresh OLVF in previously difficult cases to visually evaluate with radiography makes it possible to more accurately determine the indication for further examination using MRI, regardless of the experience or specialty of the attending physician. In addition, the evaluation of the important areas for classification will deepen our understanding of the CNN classification process, improve its reliability, and may provide new knowledge for diagnosis with the human eye

in clinical practice. This study, therefore, aims to evaluate our method to automatically determine the presence of OLVF and classify old and fresh OLVF using a CNN model with radiographs, and the areas of interest to our CNN model based on Grad-CAM images.

2. Materials and Methods

2.1. Overview of this study

The overview of this study is shown in Figure 1. In this study, we first automatically detected each lumbar vertebra from lateral radiographs. Then, after preliminary image processing, each vertebra was classified into normal, old, or fresh OLVF using three CNNs. Accuracy evaluation using external data was also performed for both (detection and classification) DL methods.

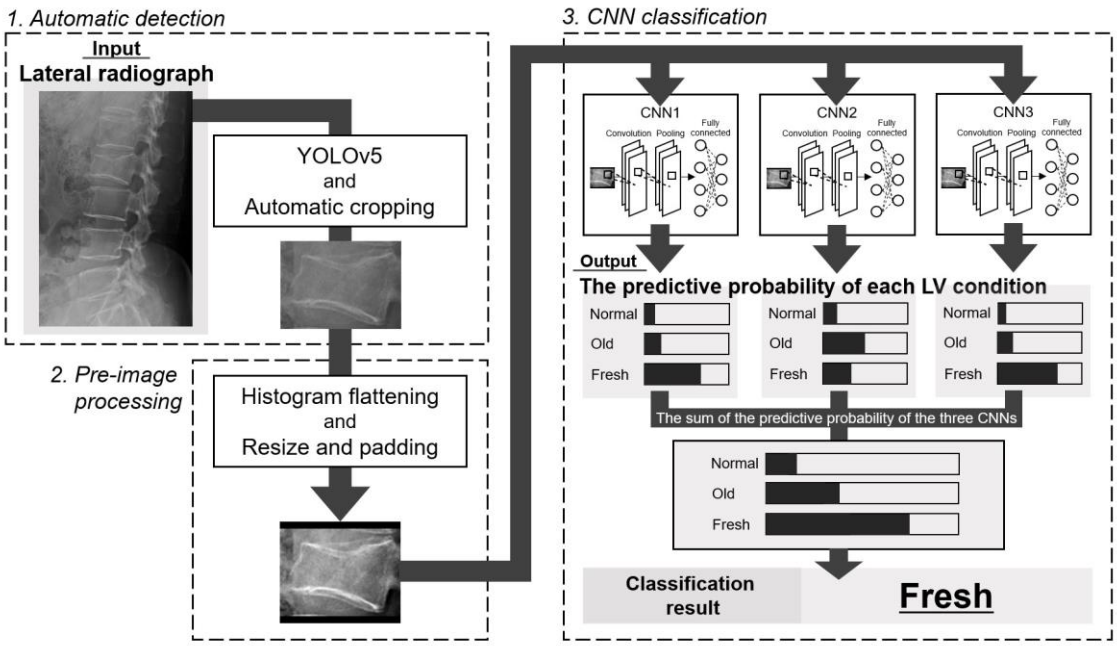

Figure 1. Overview of this study. LV means lumber vertebrae.

Normal, normal vertebra; Old, old osteoporotic lumbar vertebral fractures; Fresh, fresh osteoporotic lumbar vertebral fractures

This manuscript was constructed according to the Standards for Reporting Diagnostic Accuracy Studies (STARD) 2015 guidelines [31].

2.2. Subjects

Patients who underwent both lumbar vertebrae radiography and MRI were included in this study. Radiographs for our DL method were collected at two institutions (Institutions 1 and 2). Institutional review boards approved this study in both institutions (institution 1: No. 20-00457, September 2020, and institution 2: No. 020-0342, March 2021). Informed consent in this retrospective study was obtained from all subjects by the opt-out

method.

We collected lateral lumbar vertebral radiographs. In Institution 1, if anterior and posterior flexion imaging were acquired in addition to lateral lumbar vertebrae radiography, both images were also included. They were used as sample images to detect each lumbar vertebra automatically. Each lumbar vertebra image, after automatic cropping from lateral radiographs, was used as a sample image for CNN classification. Furthermore, in patients with thoracic vertebrae imaging, lumbar vertebrae included in the lateral thoracic vertebrae radiographs were also used for CNN classification.

In Institution 1, 523 consecutive patients with suspected OLVF who underwent radiography and MRI from March 2010 to December 2021 were included. In patients with fresh OLVF, lumbar vertebrae MRI was performed with a mean of 3.7 ± 17.0 days after radiography. Each vertebra (the first to the fifth lumbar vertebrae) was blinded to patient information and independently visually evaluated by two radiologists (14 and 12 years of experience) and classified as normal, old, or fresh OLVF. When the evaluation by two radiologists did not agree, the classification group was determined by consensus. In addition, for lumbar vertebrae that were determined to be fresh, the radiologists evaluated whether they were OLVF or pathological fractures. Pathological fractures are those resulting from bone weakness caused by primary or metastatic bone tumors.

Ninety-three patients whose visual evaluation showed that all the lumbar vertebrae from the first to the fifth were normal were excluded from this study because they may be different in the presence or absence of osteoporosis, that is, in the background of bone density, compared to patients with OLVF. To further improve the accuracy of the determination of freshness, the date of injury onset was confirmed for all patients with OLVF judged fresh in the visual evaluation. Of the 430 patients with fresh or old OLVF

in one or more vertebrae, 12 fresh and three old OLVF patients were excluded due to exclusion reasons such as severe crush, foreign substance, poor positioning, poor image quality, injury onset date unknown, or only pathological fracture (if there was OLVF other than pathological fractures, those patients were included). In this study, the criterion for severe crush was a more than 40% reduction in post-fracture vertebral body height compared to pre-fracture vertebral body height, based on Genant's criteria [32].

Exclusion reason 1 focused only on the condition of the fractured vertebrae in radiographs, while exclusion reason 2 covered all images of each vertebra after cropping from radiographs. As a result, 415 subjects were employed in this study (Figure 2a).

In Institution 2, 140 patients who underwent radiography and MRI from January 2011 to December 2021 and were diagnosed with OLVF in MRI interpretation reports by radiologists in daily practice were included in this study. In patients with fresh OLVF, lumbar vertebrae MRI was performed with a mean of 8.1 ± 10.9 days after radiography was acquired. After image collection, two fresh and one old OLVF patient were excluded based on the same exclusion criteria as in Institution 1. As a result, 137 subjects were employed in this study (Figure 2b).

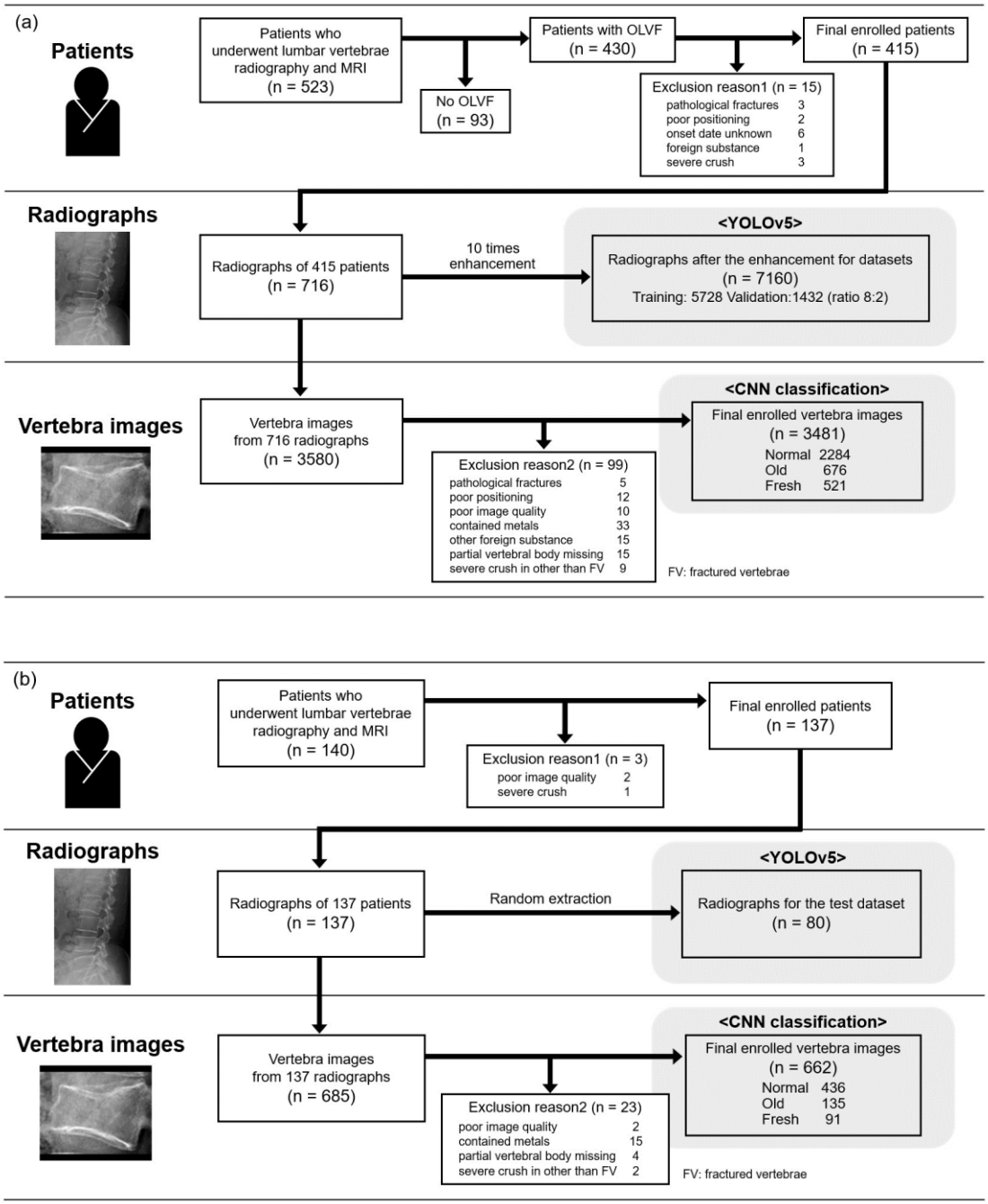

Figure 2. The number of subjects in Institutions 1 (a) and 2 (b).

Normal, normal vertebra; Old, old osteoporotic lumbar vertebral fractures; Fresh, fresh osteoporotic lumbar vertebral fractures

2.3. Image Acquisition

In Institution 1, radiographs were acquired using either the flat panel detector (FPD) of the CALNEO Smart C77 or the CALNEO MT (FUJIFILM Medical Co., Ltd., Tokyo, Japan). CALNEO Smart C77 uses CsI scintillators, and CALNEO MT uses GOS scintillators. For both FPDs, a real grid (8:1 grid ratio) (Mitaya Manufacturing Co., Ltd., Saitama, Japan) was used for scattered radiation removal instead of a scattered radiation correction process such as a virtual grid (FUJIFILM Medical Co., Ltd., Tokyo, Japan). The acquired image size was 10 × 12 inches, the pixel size was 0.15 mm, and the grayscale depth was 14 bits. The source-to-image receptor distance (SID) was 110 cm, the tube voltage was 85 kV, and the current value was automatically determined by the auto exposure control (AEC) system according to the patient's body thickness. The X-ray generator was RAD Speed Pro (SHIMADZU Corporation, Kyoto, Japan). All MRI images were acquired using a 1.5 Tesla MRI system from Ingenia (Philips Healthcare, Best, The Netherlands).

In Institution 2, radiographs were acquired using either of three FPDs. The X-ray generator and FPD in each room are shown in Table 1. The scattered radiation removal was performed on a real grid with a grid ratio of 8:1 or 10:1. The acquired image size was 14 × 17 inches, the pixel size was 0.15 mm, and the grayscale depth was 14 bits. The SID was 130 cm, the tube voltage was 90 kV, and the AEC system automatically determined the current value. An MRI was performed in a total of five rooms. The imaging equipment and magnetic field strength in each room are shown in Table 2.

Table 1. Radiography imaging equipment used in institution 2.

	Examination room1	Examination room2	Examination room3
X-ray generator	BENEO FUJIFILM Medical Corporation	RAD speed Pro SHIMADZU Corporation	Radnext 80 FUJIFILM Healthcare Corporation
X-ray detector (scintillator)	BENEO(a-Se) FUJIFILM Medical Corporation	CALNEO MT(GOS) FUJIFILM Medical Corporation	CALNEO C 1717 wireless SQ (CsI) FUJIFILM Medical Corporation
Scattered Radiation Correction	Real Grid	Real Grid	Real Grid
Grid ratio	10:1 Mitaya Corporation	8:1 Mitaya Corporation	8:1 Mitaya Corporation

Table 2. MRI imaging equipment and magnetic field strength used in institution 2.

	Examination room1	Examination room2	Examination room3	Examination room4	Examination room5
MRI system	MAGNETOM Avanto SIEMENS Healthcare	Discovery MR750w 3.0T GE Healthcare	TRILLIUM OVAL FUJIFILM Healthcare	Achieva dStream Philips Healthcare	Ingenia Elition 3.0T Philips Healthcare
Magnetic field strength (T)	1.5	3.0	3.0	1.5	3.0

2.4. Deep Learning Techniques

Three deep learning techniques were used in this study. The first is automatic lumbar vertebrae detection and cropping. The second is the classification of vertebral fracture conditions using CNN. The third is Grad-CAM image analysis.

2.4.1. Vertebral Body Detection with You Only Look Once

The automatic object detection algorithm used in this study was YOLOv5. It was trained to recognize each lumber vertebral body. There are five main models available to the public: YOLOv5 n/s/m/l/x. The main differences between versions are the automatic detection accuracy and calculation load. In this study, YOLOv5x (the largest) was selected and finetuned in training. Training and validation were conducted using 5728 training images and 1432 validation images (8:2 ratio), with a 10-fold augmentation of 716 radiographs from 415 patients in Institution 1. The image augmentation was performed using Imgaug, a Python library. Details of the image augmentation process are shown in Table 3. A total of six image processing steps were combined to create the processed image. The intensity of each process was randomly determined between the maximum and minimum values. Eighty radiographs were randomly selected out of 137 radiographs from 137 patients in Institution 2 and used for the test. One radiological technologist manually set the ground truth bounding box using the free software labelImg. All sample images were converted to 8-bit PNG images of 640 × 640 pixels. The following parameters were determined by hyperparameter evolution, a method of hyperparameter optimization using a genetic algorithm included in the YOLOv5 system: epochs, 300; batch size, 4; initial learning rate, 0.00967; momentum, 0.92755; weight decay, 0.00057.

Table 3. Details of the image augmentation process by imgaug.

	Parameter Range	
	Min	**Max**
Dropout	0.0	0.1
Gaussian Blur	0.0	0.5
Enhance Sharpness	0.0	2.5
Gamma Contrast	0.5	2.0
Gaussian Noise	0	0.05 × 255
Edge Detect	0.05	0.3

Samples and Datasets Creation

Each lumbar vertebra was automatically cropped based on the bounding box coordinates detected by YOLOv5. After cropping, a histogram flattening process was applied. The image resolution was resized and padded as necessary to 166 (W) × 140 (H) pixels (Figure 3). In this sample creation phase, 99 and 23 sample images were excluded from Institutions 1 and 2, respectively, due to the adverse conditions shown in Figure 2.

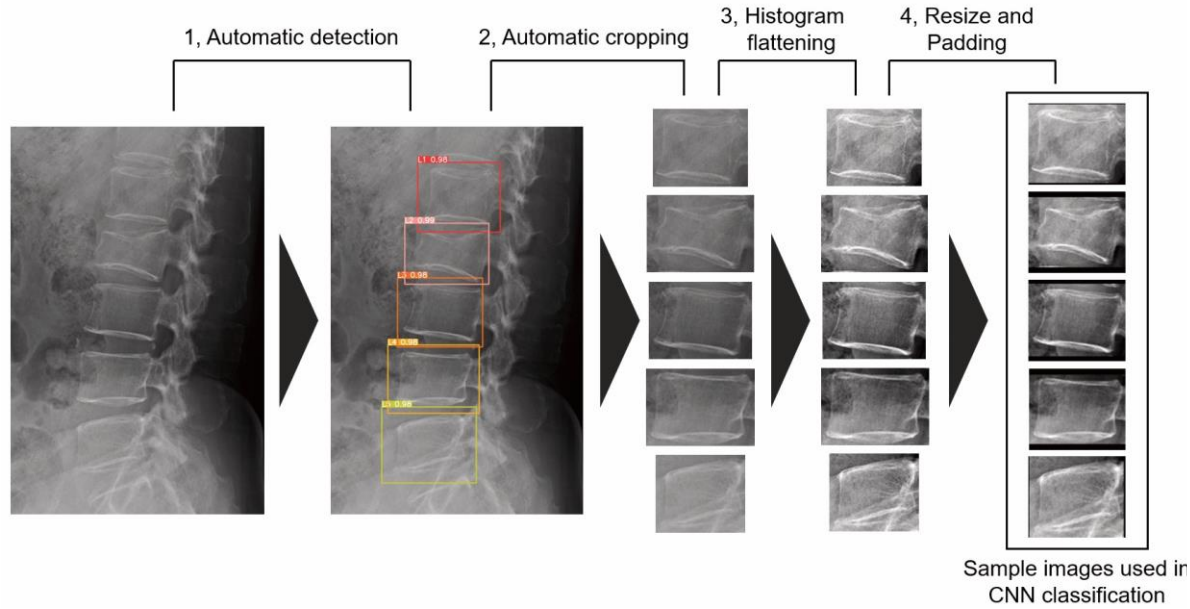

Figure 3. The flow of sample image creation. The vertebral number and the confidence value of automatic detection accompany bounding boxes detected by YOLOv5.

In Institution 1, 228, 68, and 52 sample images, or about 1/10 of the total images in the normal, old, and fresh OLVF groups, were divided, and the number of divided old and fresh OLVF images was tripled and quadrupled to resolve the imbalance in the number of images among each group. As a result, 228, 204, and 208 sample images were prepared in the normal, old, and fresh OLVF groups, respectively, as the test dataset images. After

the test dataset image division, a total of 6833 sample images (2056 normal, 2432 old, and 2345 fresh sample images) were divided in the ratio of training 8: validation 2. In Institution 2, no augmentation of the number of sample images was performed. As a result, 436, 135, and 91 sample images were prepared in the normal, old, and fresh OLVF groups, respectively. The sample images in Institution 2 were used for external validation (Figure 4).

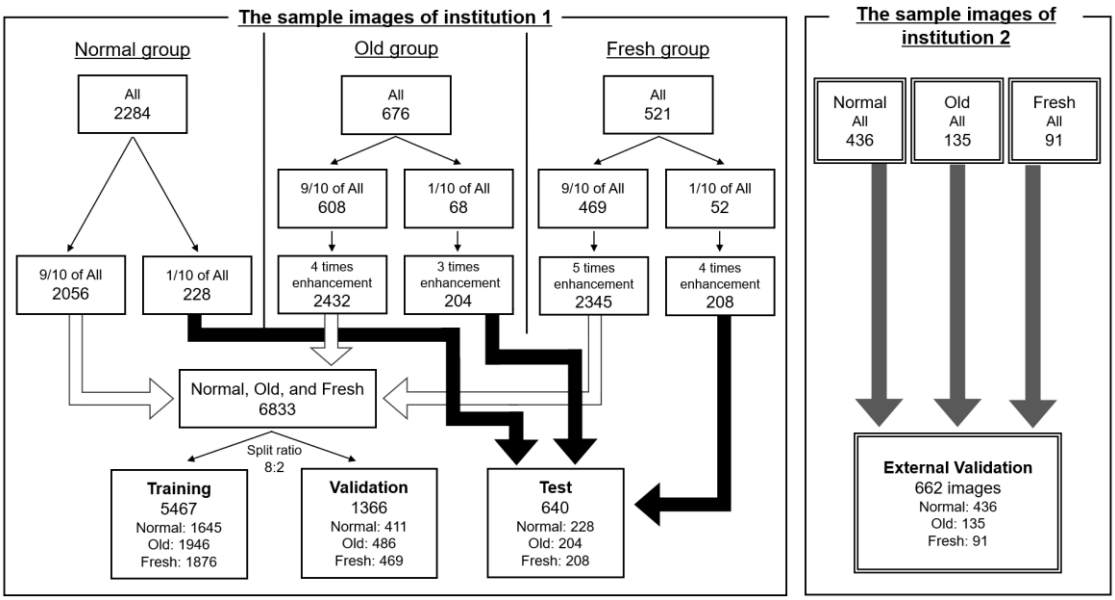

Figure 4. Datasets creation for the CNN classification.

Normal, normal vertebra; Old, old osteoporotic lumbar vertebral fractures; Fresh, fresh osteoporotic lumbar vertebral fractures

A total of four datasets were prepared: training, validation, test, and external validation. The training datasets were used to create the model to automatically determine the presence of OLVF and classify old and fresh OLVF, while the validation datasets were used to adjust the hyperparameters. The test dataset was used to evaluate the classification

performance by using images with the same characteristics as those used for training and validation, while external validation was an evaluation of the classification performance on completely unknown images. The important point is that the images in the test and external validation datasets were not used for training or validation. The robustness of the model created in this study was evaluated in more detail by also performing external validation. The parameter settings were the same as in YOLOv5 training (Table 3). Examples of processed images are shown in Figure 5.

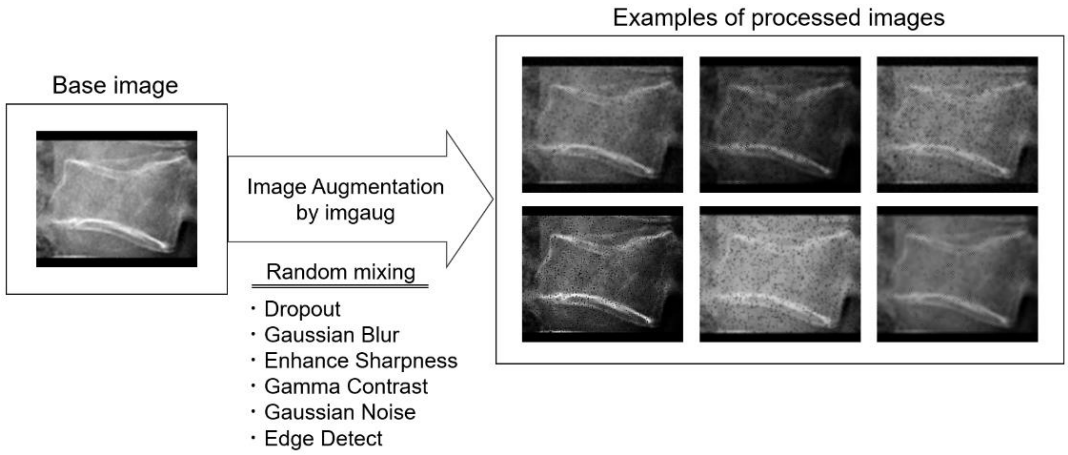

Figure 5. Examples of processed images by imgaug. A total of six image processing steps were combined to create the processed image.

2.4.2. CNN Classification

In this study, a 3-class CNN classification was performed. All sample images are classified into normal, old, or fresh OLVF groups. The CNNs output the probability that the input image is normal, old, or fresh OLVF. An ensemble model using three CNNs was employed in this study. An ensemble approach was used in which the predictive probabilities output by the three CNNs were summed for each group. The class with the

highest sum of the predictive probabilities output by three CNNs for each classification group was determined as the result of the CNN classification. The CNNs used were Resnet-50, DenseNet-161, and Res-NeXt-50. Each CNN was pre-trained with initial weights trained on ImageNet, a large image dataset on Neural Network Console (NNC) (Sony Network Communications Inc., Tokyo, Japan). Each pre-trained CNN model is available at https://nnabla.readthedocs.io/en/latest/python/api/models/imagenet.html (accessed on 15 May 2021). Before the training process, three layers were inserted just below the input layer in each CNN. The first is the "Broadcast" layer to change the color channel of input images from 1 to 3. This allows the use of grayscale images for input in CNN trained with color images. The second is the "MulScalarX" layer to divide pixel values by 255 to normalize pixel values. The third is the "ImageAugmentation" layer to pseudo-enhance the number of sample images to reduce overfitting due to the insufficient number of images. This image augmentation process included scaling, rotation, brightness, and contrast changes. The details of each enhancement process are shown in Table 4. The learning rates of Resnet-50, DenseNet-161, and ResNeXt-50 were set to 0.01, 0.001, and 0.01, respectively. The following training parameters were common to all three CNNs: 100 epochs; batch size, 4; optimizer, Nesterov.

Table 4. Details of the image enhancement process by the ImageAugmentation layer on NNC.

	Parameter Range	
	Min	Max
Image scaling	0.8	1.05
Image rotation	-20°	+20°
Brightness change	-0.2	+0.2
Contrast change	1/1.5	1.5

2.4.3. Gradient-weighted Class Activation Mapping

In Grad-CAM images, the areas of interest to the CNN are represented in redder colors. In this study, Grad-CAM images were acquired using the plug-in function of NNC. The dataset and CNN model used to create the Grad-CAM images were the test dataset and the 5th classification result in 5-fold cross-validation of DenseNet-161, which had the highest classification accuracy among all classification results, respectively. Each Grad-CAM image was divided into three groups: groups 0, 1, and 2. Groups 0, 1, and 2 contain images classified as normal vertebra, old, and fresh OLVF images by CNN classification, respectively. For example, group 0 contains correctly classified normal vertebra images, as well as old and fresh OLVF images misclassified as normal. The Grad-CAM images were then divided into 6 rows and 8 columns, for a total of 48 sections (Figure 6a), and the red element of each divided Grad-CAM image was extracted. The pixel values of the red element images were measured (Figure 6b). Each location of divided Grad-CAM images was numbered from 1 to 48 as shown in Figure 6c.

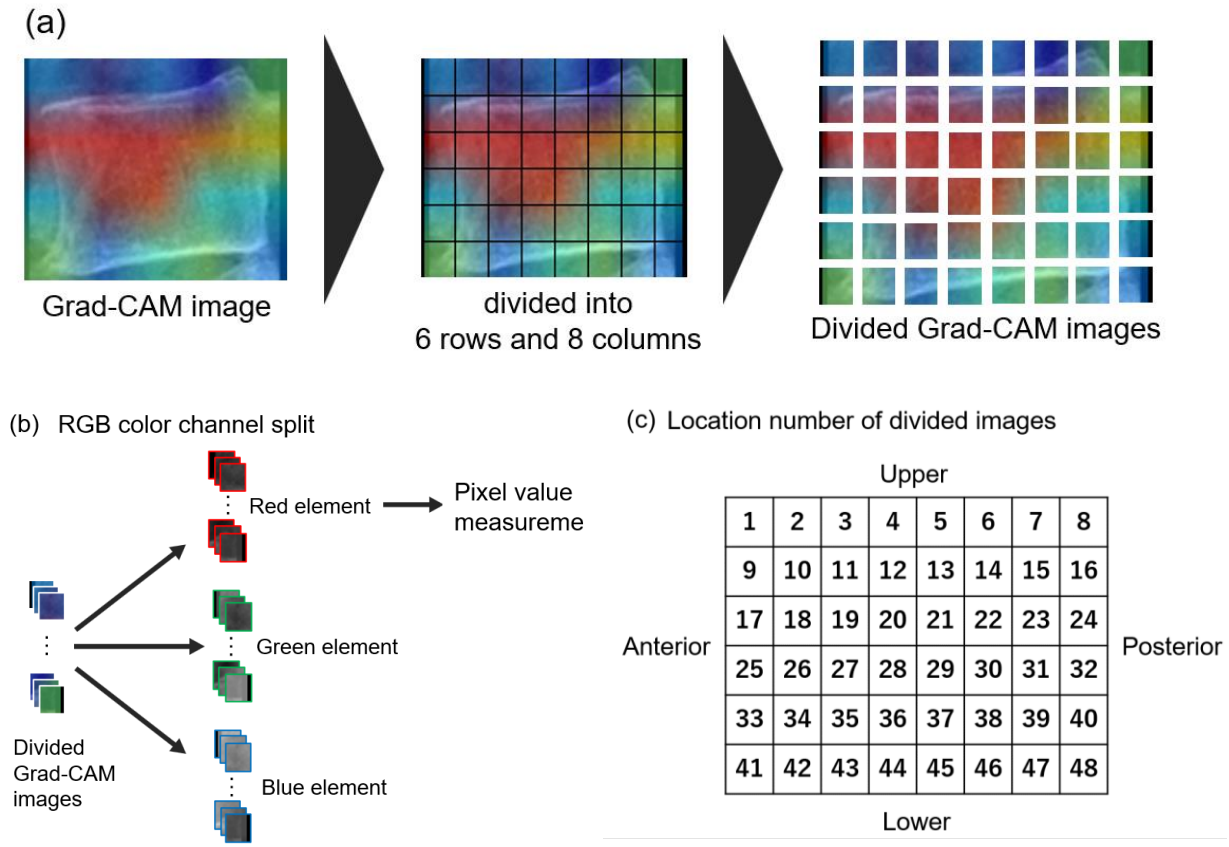

Figure 6. (a) The Grad-CAM images were then divided into 6 rows and 8 columns. (b) RGB color channel split and the pixel values measurement. (c) Location number of divided images.

2.5. Devices and Software for Deep Learning Techniques

The development environment for DL in this study was Windows 10 Pro (64bit), Intel® Core™ i7-10700KF, NVIDIA® GeForce RTX™ 3080, Python version 3.8.8, PyTorch version 1.8.1, and NNC version 2.1.0.

2.6. Statistical Analysis

Quantitative data were expressed as the mean and standard deviation. Interobserver and intraobserver visual assessment for the vertebrae condition classification were evaluated by weighted kappa values using the Landis and Koch criteria (0.0–0.2: slight agreement, 0.21–0.40: fair agreement, 0.41–0.60: moderate agreement, 0.61–0.80: substantial agreement, 0.81–1.0: almost perfect agreement) [33]. The second visual evaluation for intraobserver calculation was performed more than one month after the first assessment, and radiographs of 50 randomly selected patients from a total of 523 patients were used. All weighted kappa values were calculated using R (version 4.2.0) and the package "irr" (ver0.84.1).

The detection accuracy of the YOLOv5 model developed in this study was evaluated by the mean average precision (mAP). The mAP is the score representing the degree of agreement between the coordinates of the detected bounding box and the ground truth bounding box and is defined as follows:

$$\mathrm{mAP} = \frac{1}{\mathrm{n}} \sum_{\mathrm{k}=1}^{\mathrm{k}=\mathrm{n}} \mathrm{AP_k}$$

APk is the average precision of class k and n represents the number of classes. In this study, the mAP (0.5) and the mAP (0.5: 0.95) were used as evaluation indices. The mAP (0.5) is the mAP when the intersection over union (IoU) is set to 0.5, and the mAP (0.5: 0.95) is the average of the mAP obtained by changing IoU from 0.5 to 0.95 in 0.05 steps. IoU is an index indicating the overlap degree between the detected bounding box by YOLO and the ground truth bounding box and is calculated by dividing the common part of the two regions by the sum set.

For CNN classification, we evaluated the classification performance using 5-fold cross-validation. In this method, the datasets are divided into five groups, one of which is the validation data, and the remaining four groups are the training data, and the classification performance is evaluated. All five groups are assigned to the validation data one at a time. Training and validation were performed five times in each CNN per cross-validation component, for a total of 15 results.

The accuracy, sensitivity, specificity, false positive rate, and false negative rate were calculated for the classification performance. These classification performances were calculated by considering either normal, old, or fresh OLVF as positive and the others as negative. The name of the group considered positive was appended to each classification performance. The overall classification performance is the average of the values calculated when each group is considered positive. For example, when normal is considered positive, old and fresh OLVF are considered negative, and the accuracy is shown as $accuracy_{normal}$. $Accuracy_{all}$ is the average of $accuracy_{normal}$, $accuracy_{old}$, and $accuracy_{fresh}$.

In addition, receiver operating characteristic (ROC) curves were plotted, and the area under the curve (AUC) values were calculated. They were plotted and calculated by scikit-learn, one of the Python libraries.

A total of 95% confidence intervals (CI) for the classification performance and AUC were calculated using scikit-learn and statsmodels in the Python libraries, respectively.

The Kruskal-Wallis test and the Steel-Dwass test were performed in the assessment of the pixel values of the red element in divided Grad-CAM images, to generate the significant deference and the assessment of the group which CNN had higher interest, respectively. $P < 0.05$ was considered to indicate a statistically significant difference. The

boundary of the areas with significant differences between the two groups in each combination was determined by the support vector machine (SVM) method using scikit-learn.

3. Results

In Institution 1, a total of 716 lateral lumber vertebrae radiographs of 415 OLVF patients were employed. The subjects consisted of 280 fresh OLVF patients with a mean age of 78.5 ± 11.4 years and 135 old OLVF patients with a mean age of 77.1 ± 9.3 years. No significant difference in mean age between fresh and old OLVF patients was observed ($p > 0.05$). In 280 fresh OLVF patients, radiography in 200 patients (71.4%) was performed within 14 days from the injury onset.

In Institution 2, a total of 137 lateral lumber vertebrae radiographs of 137 OLVF patients were employed. The subjects consisted of 77 fresh OLVF patients with a mean age of 69.6 ± 13.7 years and 60 old OLVF patients with a mean age of 70.8 ± 11.1 years. No significant difference in mean age between fresh and old OLVF patients was observed ($p > 0.05$). In 77 fresh OLVF patients, radiography was performed in 48 patients (62.3%) within 14 days from the injury onset.

3.1. Agreement Rate in the Visual Evaluation of Each Vertebral Body

The interobserver agreement value for visual evaluation by two radiologists was 0.801. The intraobserver agreement values for raters 1 and 2 were 0.821 and 0.861, respectively. The agreement in all evaluations was almost perfect. Consensus was required in 141 of 523 cases because of inconsistent evaluation of at least one vertebra.

3.2. YOLOv5

The detection performance of YOLOv5 was mAP (0.5) of 0.995 and mAP (0.5: 0.95) of 0.993 for the validation dataset and mAP (0.5) of 0.982 and mAP (0.5: 0.95) of 0.835 for the test dataset in terms of detection for lumber vertebrae. Using automatic cropping based on the bounding box detected by YOLOv5, vertebra images of 2284 normal, 676 old, and 521 fresh OLVF were produced in Institution 1. Similarly, vertebra images of 436 normal, 135 old, and 91 fresh OLVF were produced in Institution 2. The breakdown of vertebral body numbers is shown in Table 5.

Table 5. The breakdown of each vertebral body number in this study.

	Institution 1			**Institution 2**		
	Normal	**Fresh**	**Old**	**Normal**	**Fresh**	**Old**
The number of all vertebra images	2284	521	676	436	91	135
L1 (%)	322 (14.1)	175 (33.6)	211 (31.2)	63 (14.4)	26 (28.6)	43 (31.9)
L2 (%)	467 (20.4)	118 (22.6)	124 (18.3)	83 (19.0)	23 (25.3)	27 (20.0)
L3 (%)	464 (20.3)	106 (20.3)	131 (19.4)	87 (20.0)	19 (20.9)	28 (20.7)
L4 (%)	486 (21.3)	76 (14.6)	129 (19.1)	97 (22.2)	15 (17.0)	23 (17.0)
L5 (%)	545 (23.9)	46 (8.8)	81 (12.0)	106 (24.3)	8 (8.8)	14 (10.4)

Normal, normal vertebra; Old, old osteoporotic lumbar vertebral fractures; Fresh, fresh osteoporotic lumbar vertebral fractures

3.3. Classification Performance by CNNs

The confusion matrix in this classification is shown in Figure 7. The classification performance was as follows: The $accuracy_{all}$, $sensitivity_{all,}$ $specificity_{all}$, false positive $rate_{all}$, false negative $rate_{all}$, and AUC_{all} was 0.894 [CI: 0.870–0.917], 0.836 [CI: 0.790–0.882], 0.920 [CI: 0.894–0.946], 0.161 [CI: 0.111–0.211], 0.077 [CI: 0.053–0.100], and 0.912 [CI: 0.861–0.963] in the test dataset, and 0.867 [CI: 0.841–0.892], 0.674 [CI:

0.603–0.743], 0.866 [CI: 0.832–0.898], 0.288 [CI: 0.217–0.363], 0.111 [CI: 0.079–0.144], and 0.855 [CI: 0.767–0.943] in the external validation dataset, respectively. A high $accuracy_{all}$ of 0.86 or higher in both datasets were achieved. The ROC curves for each dataset are shown in Figure 8.

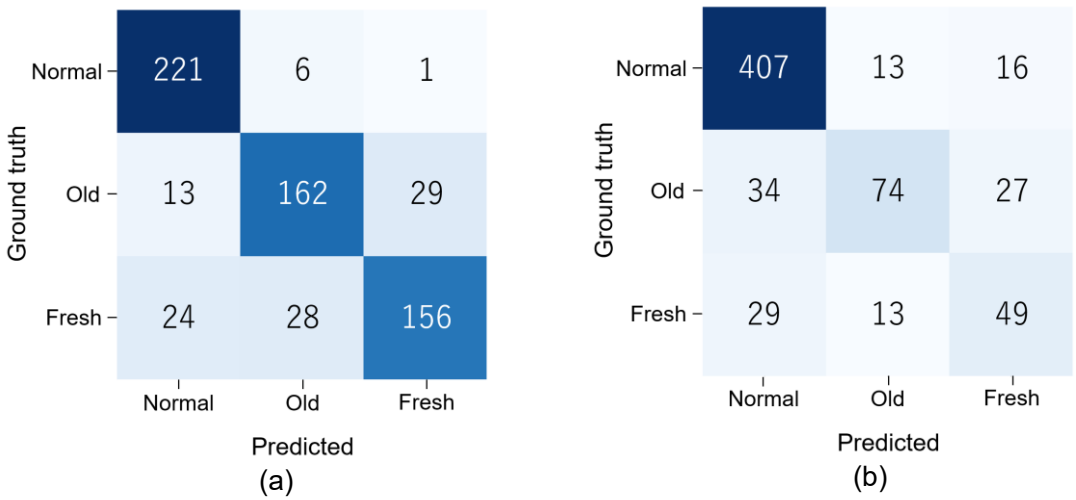

Figure 7. The confusion matrix in CNN classification in the test (a) and external validation (b) datasets.

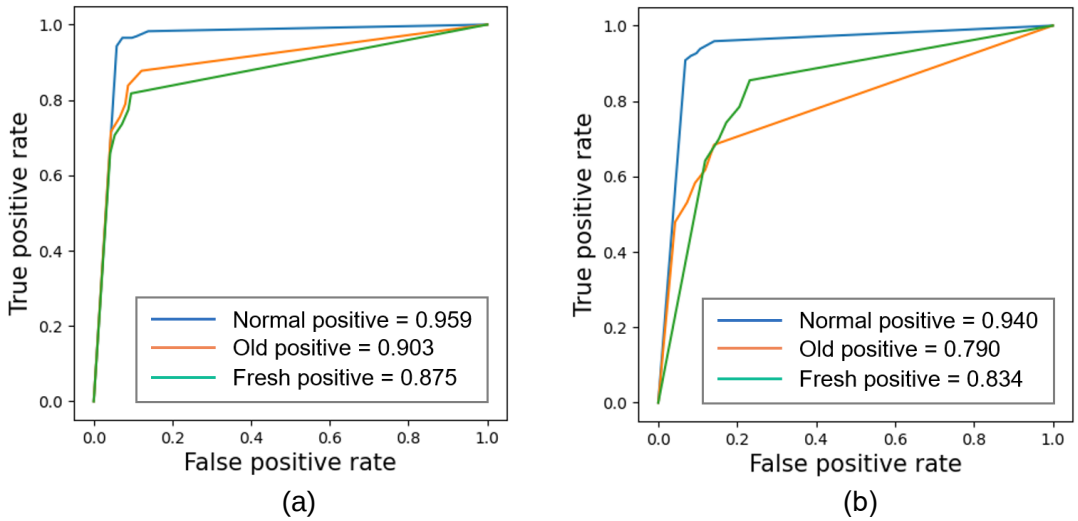

Figure 8. ROC curve for the test (a) and external validation (b) datasets.

Normal, normal vertebra; Old, old osteoporotic lumbar vertebral fractures; Fresh, fresh osteoporotic lumbar vertebral fractures

The classification performances in both datasets were summarized in Table 6.

Table 6. The summary of each classification performance.

	Test				**External Validation**			
	The Group Considered Positive				**The Group Considered Positive**			
	All	**Normal**	**Old**	**Fresh**	**All**	**Normal**	**Old**	**Fresh**
Accuracy	0.894	0.930	0.880	0.871	0.867	0.861	0.868	0.872
Sensitivity	0.836	0.968	0.791	0.748	0.674	0.933	0.550	0.538
Specificity	0.920	0.908	0.922	0.930	0.866	0.722	0.950	0.925
False positive rate	0.161	0.146	0.175	0.163	0.288	0.134	0.264	0.465
False negative rate	0.077	0.019	0.096	0.115	0.111	0.152	0.108	0.074
AUC	0.912	0.959	0.903	0.875	0.855	0.940	0.790	0.834

Normal, normal vertebra; Old, old osteoporotic lumbar vertebral fractures; Fresh, fresh osteoporotic lumbar vertebral fractures

3.4. Grad-CAM image analysis

The average pixel value in each group is shown in Figure 9 and Table 7. Focusing on the column-by-column variation, Grad-CAM images of all groups had higher pixel values around the center of the image (columns 3, 4, 5, and 6) and lower pixel values in the outer parts of the image (columns 1, 2, 7, and 8). Focusing on the row-by-row variation, the three highest pixel values were in rows 4, 5, and 6 in the normal (0) group, rows 2, 3, and 4 in the old (1) group, and rows 3, 4, and 5 in the fresh (2) group, respectively.

The results of the Kruskal-Wallis test and the Steel-Dwass test are shown in Figure 10. The sections where significant differences occurred were shown in color. In addition, groups with significantly higher values were indicated with the group number: 0, 1, or 2.

There was a definite difference in the section with high interest in each group.

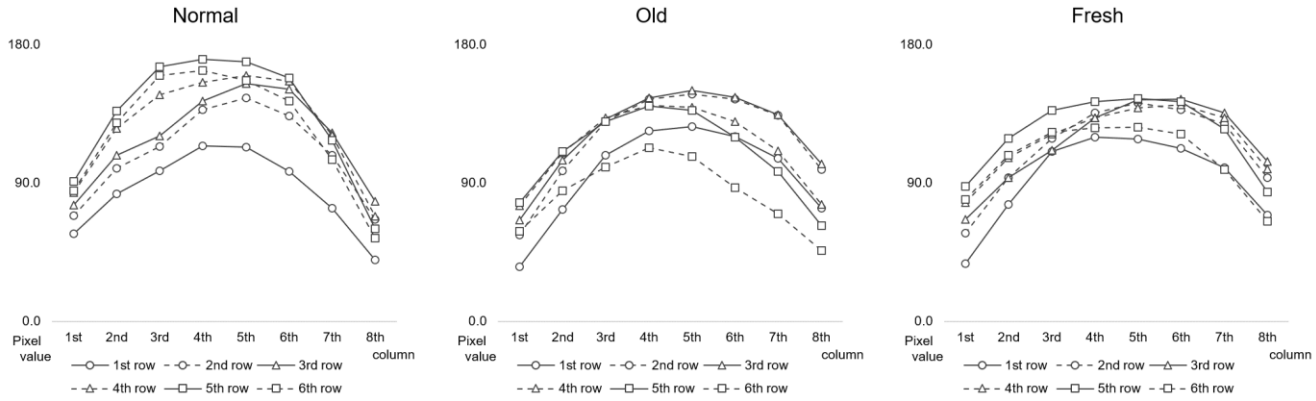

Figure 9. The average pixel value of each red element image per row.

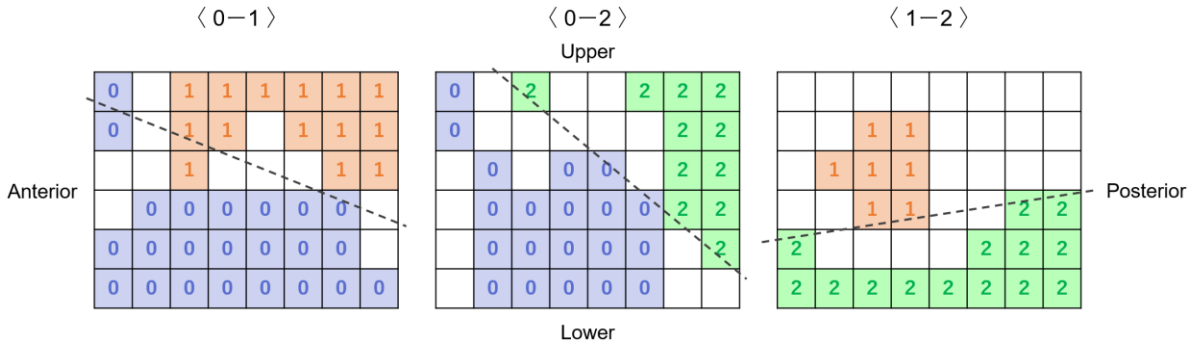

Figure 10. The results of the Kruskal-Wallis test and the Steel-Dwass test.

Significant differences in the Kruskal-Wallis test were found in the colored section. Groups with significantly higher values were indicated with the group number: 0, 1, or 2. Groups 0, 1, and 2 contain images classified as normal vertebra, old, and fresh OLVF images by CNN classification, respectively. Dotted lines indicate the boundary of the areas with significant differences between the two groups in each combination determined by the SVM method.

Table 7. The average pixel value of each red element image.

Average pixel value							
Location number	**Normal**	**Old**	**Fresh**	Location number	**Normal**	**Old**	**Fresh**
1	57.22	35.50	37.66	25	83.68	75.41	77.55
2	83.00	73.03	76.09	26	125.61	109.69	106.55
3	97.84	108.34	110.42	27	147.62	132.44	121.83
4	114.15	123.80	119.80	28	155.54	140.43	132.39
5	113.37	126.71	118.82	29	159.93	139.19	138.97
6	97.72	120.30	112.63	30	156.51	129.97	140.97
7	73.87	106.22	100.10	31	121.20	110.96	132.58
8	40.06	73.89	69.26	32	68.28	76.14	99.06
9	69.01	56.41	57.38	33	91.18	77.22	87.90
10	99.45	98.06	93.38	34	136.85	110.64	119.21
11	113.73	129.91	118.93	35	165.63	130.00	137.21
12	137.56	144.51	135.55	36	170.52	140.38	143.04
13	145.44	148.07	142.38	37	169.08	137.32	145.04
14	133.70	144.58	137.77	38	158.54	120.00	143.16
15	108.06	134.21	128.35	39	117.78	97.56	125.02
16	66.56	98.81	93.39	40	60.41	62.45	84.32
17	75.80	65.91	66.39	41	85.01	58.78	79.42
18	108.04	105.10	93.76	42	129.17	85.14	108.32
19	120.92	131.52	111.57	43	160.10	100.58	123.25
20	143.47	145.56	132.30	44	163.13	112.86	126.14
21	154.75	150.45	144.08	45	156.72	107.40	126.28
22	150.97	146.05	144.54	46	143.37	87.20	121.99
23	122.78	134.38	135.91	47	105.36	70.01	98.86
24	77.99	102.59	104.12	48	54.27	46.19	65.06

Normal, normal vertebra; Old, old osteoporotic lumbar vertebral fractures; Fresh, fresh osteoporotic lumbar vertebral fractures

4. Discussion

In this retrospective study, we attempted to develop an automatic method to detect OLVF and classify old and fresh OLVF by creating a CNN model with lateral lumbar vertebrae radiographs. This is the first study to validate the generalization performance of 3-class OLVF classification using images from multiple facilities. Some images in the external

validation dataset differed from those used in the training and validation regarding imaging conditions that affect image quality, such as the tube voltage and grid ratio. The fact that our method is effective even for such images may suggest that this model is quite versatile.

Objectivity must be ensured in the sample creation process. In this study, YOLOv5 was applied to the sample creation process. This eliminates human bias caused by manual procedures and reduces sample creation time. To the best of our knowledge, though one study attempted to identify fresh vertebral fractures on radiography using a CNN, that study had the limitation that the ROIs to extract the target vertebrae were manually drawn [30], which has been overcome in our study.

Since Grad-CAM images are generally evaluated visually, subjective differences among observers occur, and it is difficult to find a specific trend visually from many sample images. To solve this problem, we attempted to quantify Grad-CAM images. First, we found that the CNN model focuses mainly on the center of the sample images to classify the conditions of the vertebral fracture. In this study, padding and image augmentation of angle change were used to uniform the sample image size and to increase the number of images, respectively. Therefore, the periphery of the sample images includes areas with a pixel value of 0 (Figure 11). Although there was concern that these areas would harm image classification by CNN, the vertebral body part in the sample images was of relatively high interest in this study, and it was considered that the CNN classification was correctly based on the characteristics of the vertebral body rather than on background areas such as the soft tissue.

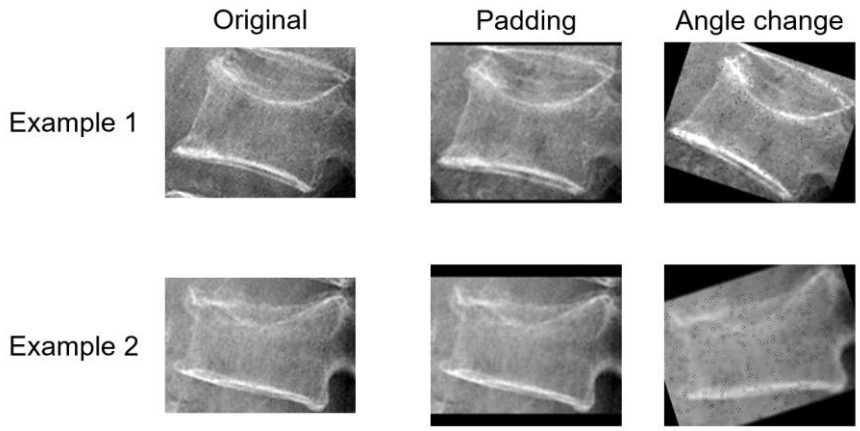

Figure 11. An example of the sample images includes areas with a pixel value of 0 (Padding, Angle change). Black areas indicate those with a pixel value of 0.

Second, there was a definite difference in the section with high interest in each group. OLVF can be classified into three categories: wedge, biconcave, and crush. Most of them are wedge and biconcave fractures, which are the most common and the second-most common, respectively. The wedge fractures are characterized by compression of the anterior wall of the vertebral body. In the biconcave fractures, only the middle portion of the vertebral body is collapsed, whereas the anterior and posterior walls remain intact [34]. From the above, the vertebral body shape is maintained in normal vertebrae, whereas the anterior and middle portions may often collapse in many OLVF vertebrae. In a comparison between old and fresh groups, the line of the upper endplate that has fallen to the vertebral body part in old OLVF, and the line of the lower endplate that retain its shape in fresh OLVF were features CNN focused on for classification. These suggests that the area where the shape of vertebral body, especially the anterior vertebral wall and endplates, differs significantly from that of the comparison group is the point of interest for the CNN. Nevertheless, because the only degree of vertebral height reduction does not

unconditionally allow to classify vertebrae conditions, the CNN may have uniquely determined an additional important factor for correct classification that is not recognized by the human eye, in addition to the feature of the anterior vertebral wall and endplates. In addition to areas with large deformations that appear more distinctive to humans, areas with smaller deformations were also considered to be important features for the CNN. The slope of the boundary between the normal vertebrae and the fresh OLVF group was greater than that between the normal vertebrae and the old OLVF group, possibly reflecting the effect of physiological lordosis because a higher proportion of upper lumbar vertebrae in the fresh OLVF than in the old OLVF in this study.

Although there were several cases in which the correct classification was achieved even though the CNN did not focus on the areas with significant high interest section, the results acquired in this Grad-CAM image analysis may have identified areas that are statistically more noteworthy.

In this study, we used an ensemble approach in which the predictive probabilities output by the three CNNs are summed for each group (Figure 12). Some previous studies had used an ensemble model with majority voting to determine the final result [10, 35]. However, when classification targets are three groups as in this study, that method cannot determine the final result if there is only one classified result for each group. In addition, even if there is a high probability output but it is only one vote, that suggestion with the high probability output may be ignored. The method used in this study can solve these problems and determine more statistically reliable classification results.

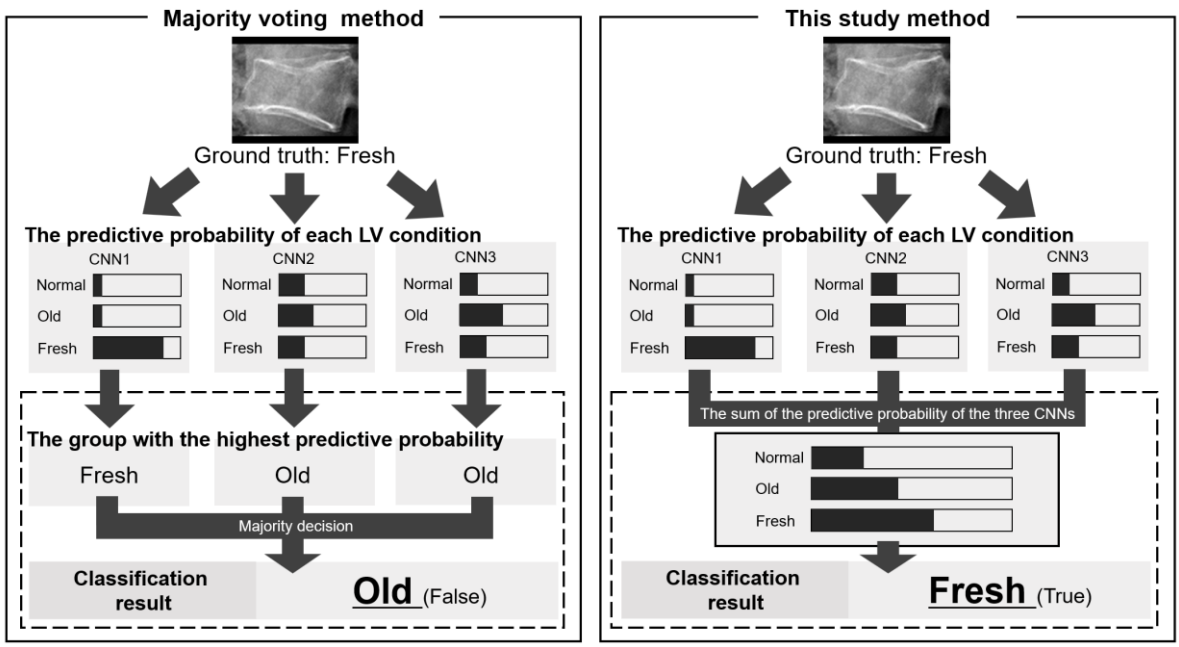

Figure 12. An ensemble approach in this study method.

A study reported by Strickland et al. (2023) [36] to determine fresh OLVF based on findings on radiography showed a sensitivity of 52% and a specificity of 95%. Langdon et al. stated that it is not possible to distinguish between fresh and old fractures on radiographs [16]. As described above, the human evaluation to determine fresh OLVF using radiographs is very difficult. The sensitivity of 84% and 67% and specificity of 92% and 87% (test/external validation) were achieved in the proposed method. Compared to the diagnostic accuracy by radiographs alone as reported by Strickland et al., the sensitivity was at least 15% higher and the specificity was comparable even in the more difficult situation, the classification in the external validation dataset. We believe that the proposed method combining CNN with radiography has high classification and generalization performance and would further improve the usefulness of radiography.

The implementation of this method may benefit both the physician and the patient. In

clinical practice, physicians carefully observe each vertebral body on radiographs to determine the diagnosis. By referring to the objective and consistent classification results provided by the CNN model developed in this study, physicians could reduce the burden associated with the evaluation of radiography and make a diagnosis and therapeutic strategy in a shorter time. In addition, the proposed method's improved fresh OLVF screening accuracy prevents missed fresh OLVF and reduces unnecessary MRI imaging. As a result, it will enable a more efficient selection of patients who require close examination by MRI. Furthermore, the proposed method requires only radiography, one of the most widely used imaging exams. This greatly benefits patients by allowing them to receive high-accuracy screening at any given facility. The proposed method might be more effective when accurate diagnosis is difficult due to the co-existence of old and fresh OLVFs on radiographs, even if the physicians are inexperienced in evaluating OLVFs.

The current challenge for clinical practice is that the target lumbar vertebrae images must be extracted from the image storage server and transferred to a device for machine learning each time.

The limitations of this study are as follows: First, cases of lumber vertebrae with deformation/crush, strong scoliosis, and metal material implantation were excluded. If the lumber vertebrae with such characteristics are input into the CNN created in this study, it may be unable to output the correct diagnosis because it was not trained on such images. Second, since this study targeted OLVF, it is unclear whether high classification accuracy can be guaranteed for pathological fractures caused by bony metastasis. It may be difficult to utilize this CNN at facilities with many pathological fracture cases. Third, the number of sample images is limited. To create a more accurate and versatile network, it is necessary to increase the number of samples, collect image data from more facilities, and

conduct training using images with various characteristics.

5. Conclusions

The proposed CNN-based method demonstrated high performance in determining the presence of OLVF and classifying old or fresh OLVF on radiography. Utilizing objective classification results from our CNN is expected to improve the accuracy of fresh OLVF screening. This may lead to appropriate decisions on the indication for close examination with MRI.

In addition, the quantitative evaluation of Grad-CAM images allowed us to identify the areas of interest for the CNN model created in this study, which were found to be mainly the anterior vertebral wall and endplates. Further detailed Grad-CAM analysis might provide new knowledge for OLVF evaluation with the human eye in clinical practice in the future.

6. References

1. Xu, W.; Wang, S.; Chen, C.; Li, Y.; Ji, Y.; Zhu, X.; Li, Z. Correlation analysis between the magnetic resonance imaging characteristics of osteoporotic vertebral compression fractures and the efficacy of percutaneous vertebroplasty: a prospective cohort study. BMC Musculoskelet Disord. 2018, 19(1), 114, doi: 10.1186/s12891-018-2040-8.
2. Su, WC.; Wu, WT.; Peng, CH.; Yu, TC.; Lee, RP.; Wang, JH.; Yeh, KT. The Short-Term Changes of the Sagittal Spinal Alignments After Acute Vertebral Compression Fracture Receiving Vertebroplasty and Their Relationship With the Change of Bathel

Index in the Elderly. Geriatr Orthop Surg Rehabil. 2022, 13, 21514593221100238, doi: 10.1177/21514593221100238.

3. Manhard, MK.; Nyman, JS.; Does, MD. Advances in imaging approaches to fracture risk evaluation. Transl Res. 2017, 181, 1-14. doi: 10.1016/j.trsl.2016.09.006.
4. Tan, E.; Wang, T.; Pelletier, MH.; Walsh, WR. Effects of cement augmentation on the mechanical stability of multilevel spine after vertebral compression fracture. J Spine Surg. 2016, 2(2), 111-121. doi: 10.21037/jss.2016.06.05.
5. Zhang, H.; Xu, C.; Zhang, T.; Gao, Z.; Zhang, T. Does Percutaneous Vertebroplasty or Balloon Kyphoplasty for Osteoporotic Vertebral Compression Fractures Increase the Incidence of New Vertebral Fractures? A Meta-Analysis. Pain Physician. 2017, 20(1), E13-E28.
6. Jin, C.; Xu, G.; Weng, D.; Xie, M.; Qian, Y. Impact of Magnetic Resonance Imaging on Treatment-Related Decision Making for Osteoporotic Vertebral Compression Fracture: A Prospective Randomized Trial. Med Sci Monit. 2018, 24, 50-57. doi: 10.12659/msm.905729.
7. Cheng, J.; Muheremu, A.; Zeng, X.; Liu, L.; Liu, Y.; Chen, Y. Percutaneous vertebroplasty vs balloon kyphoplasty in the treatment of newly onset osteoporotic vertebral compression fractures: A retrospective cohort study. Medicine (Baltimore). 2019, 98(10), e14793, doi: 10.1097/MD.0000000000014793.
8. Suzuki, N.; Ogikubo, O.; Hansson, T.; The prognosis for pain, disability, activities of daily living and quality of life after an acute osteoporotic vertebral body fracture: its relation to fracture level, type of fracture and grade of fracture deformation. Eur Spine J. 2009, 18(1), 77-88, doi: 10.1007/s00586-008-0847-y.
9. Shigenobu, K.; Hashimoto, T.; Kanayama, M.; Ohha, H.; Yamane, S. The efficacy of

osteoporotic treatment in patients with new spinal vertebral compression fracture pain, ADL, QOL, bone metabolism and fracture-healing - In comparison with weekly teriparatide with bisphosphonate. Bone Rep. 2019, 11, 100217, doi: 10.1016/j.bonr.2019.100217.

10. Li, YC.; Chen, HH.; Horng-Shing, Lu H.; Hondar, Wu HT.; Chang, MC.; Chou, PH. Can a Deep-learning Model for the Automated Detection of Vertebral Fractures Approach the Performance Level of Human Subspecialists? Clin Orthop Relat Res. 2021, 479(7), 1598-1612, doi: 10.1097/CORR.0000000000001685.
11. Lenski, M.; Büser, N.; Scherer, M. Concomitant and previous osteoporotic vertebral fractures. Acta Orthop. 2017, 88(2), 192-197, doi: 10.1080/17453674.2016.1273644. Epub 2017 Jan 6.
12. Choi, WH.; Oh, SH.; Lee, CJ.; Rhim, JK.; Chung, BS.; Hong, HJ. Usefulness of SPAIR Image, Fracture Line and the Adjacent Discs Change on Magnetic Resonance Image in the Acute Osteoporotic Compression Fracture. Korean J Spine. 2012, 9(3), 227-231, doi: 10.14245/kjs.2012.9.3.227.
13. Marongiu, G.; Congia, S.; Verona, M.; Lombardo, M.; Podda, D.; Capone, A. The impact of magnetic resonance imaging in the diagnostic and classification process of osteoporotic vertebral fractures. Injury. 2018, 49 Suppl 3, S26-S31, doi: 10.1016/j.injury.2018.10.006.
14. Lin, HH.; Chou, PH.; Wang, ST.; Yu, JK.; Chang, MC.; Liu, CL. Determination of the painful level in osteoporotic vertebral fractures--Retrospective comparison between plain film, bone scan, and magnetic resonance imaging. J Chin Med Assoc. 2015, 78(12), 714-718, doi: 10.1016/j.jcma.2015.06.015.
15. Langdon, J.; Way, A.; Heaton, S.; Bernard, J.; Molloy, S. Vertebral compression

fractures--new clinical signs to aid diagnosis. Ann R Coll Surg Engl. 2010, 92(2), 163-166, doi: 10.1308/003588410X12518836440162.

16. Bierry, G.; Venkatasamy, A.; Kremer, S.; Dosch, JC.; Dietemann, JL. Dual-energy CT in vertebral compression fractures: performance of visual and quantitative analysis for bone marrow edema demonstration with comparison to MRI. Skeletal Radiol. 2014, 43(4), 485-492, doi: 10.1007/s00256-013-1812-3.
17. Kim, DH.; Jeong, JG.; Kim, YJ.; Kim, KG.; Jeon, JY. Automated Vertebral Segmentation and Measurement of Vertebral Compression Ratio Based on Deep Learning in X-Ray Images. J Digit Imaging. 2021, 34(4), 853-861, doi: 10.1007/s10278-021-00471-0.
18. Kim, KC.; Cho, HC.; Jang, TJ.; Choi, JM.; Seo, JK. Automatic detection and segmentation of lumbar vertebrae from X-ray images for compression fracture evaluation. Comput Methods Programs Biomed. 2021, 200, 105833, doi: 10.1016/j.cmpb.2020.105833.
19. Diekhoff, T.; Engelhard, N.; Fuchs, M.; Pumberger, M.; Putzier, M.; Mews, J.; Makowski, M.; Hamm, B.; Hermann, KA. Single-source dual-energy computed tomography for the assessment of bone marrow oedema in vertebral compression fractures: a prospective diagnostic accuracy study. Eur Radiol. 2019, 29(1), 31-39, doi: 10.1007/s00330-018-5568-y.
20. Maselli, F.; Rossettini, G.; Viceconti, A.; Testa, M. Importance of screening in physical therapy: vertebral fracture of thora-columbar junction in a recreational runner. BMJ Case Rep. 2019, 12(8): e229987. doi: 10.1136/bcr-2019-229987.
21. Suzuki, N.; Ogikubo, O.; Hansson, T. Previous vertebral compression fractures add to the deterioration of the disability and quality of life after an acute compression

fracture. Eur Spine J. 2010, 19(4), 567-574. doi: 10.1007/s00586-009-1162-y.

22. Oudshoorn, C.; Hartholt, KA.; Zillikens, MC.; Panneman, MJ.; van der Velde, N.; Colin, EM.; Patka, P.; van der Cammen, TJ. Emergency department visits due to vertebral fractures in the Netherlands, 1986-2008: steep increase in the oldest old, strong association with falls. Injury. 2012, 43(4), 458-461, doi: 10.1016/j.injury.2011.09.014.
23. Cao, Y.; Yu, J.; Zhang, H.; Xiong, J.; Luo, Z. Classification of hepatic cavernous hemangioma or hepatocellular carcinoma using a convolutional neural network model. J Gastrointest Oncol. 2022, 13(2), 787-791, doi: 10.21037/jgo-22-197.
24. Halme, HL.; Ihalainen, T.; Suomalainen, O.; Loimaala, A.; Mätzke, S.; Uusitalo, V.; Sipilä, O.; Hippeläinen, E. Convolutional neural networks for detection of transthyretin amyloidosis in 2D scintigraphy images. EJNMMI Res. 2022, 12(1), 27, doi: 10.1186/s13550-022-00897-9.
25. Guo, G.; Zhang, Z. Road damage detection algorithm for improved YOLOv5. Sci Rep. 2022, 12(1), 15523, doi: 10.1038/s41598-022-19674-8.
26. Chen, S.; Duan, J.; Wang, H.; Wang, R.; Li, J.; Qi, M.; Duan, Y.; Qi, S. Automatic detection of stroke lesion from diffu-sion-weighted imaging via the improved YOLOv5. Comput Biol Med. 2022, 150, 106120, doi: 10.1016/j.compbiomed.2022.106120.
27. Wang, C.; Zhang, Y.; Zhou, Y.; Sun, S.; Zhang, H.; Wang, Y. Automatic detection of indoor occupancy based on improved YOLOv5 model. Neural Comput Appl. 2023, 35(3), 2575-2599, doi: 10.1007/s00521-022-07730-3.
28. Shi, Y.; Li, J.; Yu, Z.; Li, Y.; Hu, Y.; Wu, L. Multi-Barley Seed Detection Using iPhone Images and YOLOv5 Model. Foods. 2022, 11(21), 3531, doi:

10.3390/foods11213531.

29. R, R, Selvaraju.; M, Cogswell.; A, Das.; R, Vedantam.; D, Parikh.; D, Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE International Conference on Computer Vision. 2017, 618-626, doi: 10.1109/ICCV.2017.74.
30. Chen, W.; Liu, X.; Li, K.; Luo, Y.; Bai, S.; Wu, J.; Chen, W.; Dong, M.; Guo, D. A deep-learning model for identifying fresh vertebral compression fractures on digital radiography. Eur Radiol. 2022, 32(3), 1496-1505, doi: 10.1007/s00330-021-08247-4.
31. Bossuyt, PM.; Reitsma, JB.; Bruns, DE.; Gatsonis, CA.; Glasziou, PP.; Irwig, L.; Lijmer, JG.; Moher, D.; Rennie, D.; de Vet HC.; Kressel, HY.; Rifai, N.; Golub, RM.; Altman, DG.; Hooft, L.; Korevaar, DA.; Cohen, JF. STARD 2015: an updated list of es-sential items for reporting diagnostic accuracy studies. BMJ. 2015, 351, h5527. doi: 10.1136/bmj.h5527.
32. Genant HK.; Wu CY.; van Kuijk C.; Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res. 1993, 8(9), 1137-1148, doi:10.1002/jbmr.5650080915
33. Landis, JR.; Koch, GG. The measurement of observer agreement for categorical data. Biometrics. 1977, 33(1), 159-174.
34. Alexandru, D.; So, W. Evaluation and management of vertebral compression fractures. Perm J. 2012;16(4), 46-51, doi:10.7812/TPP/12-037
35. Li, W.; Yu, S.; Yang, R.; Tian, Y.; Zhu, T.; Liu, H.; Jiao, D.; Zhang, F.; Liu, X.; Tao, L.; Gao, Y.; Li, Q.; Zhang, J.; Guo, X. Machine Learning Model of ResNet50-Ensemble Voting for Malignant-Benign Small Pulmonary Nodule Classification on Computed Tomography Images. Cancers (Basel). 2023, 15;15(22):5417. doi:

10.3390/cancers15225417.

36. Strickland, CD.; DeWitt, PE.; Jesse, MK.; Durst, MJ.; Korf, JA. Radiographic assessment of acute vs chronic vertebral compression fractures. Emerg Radiol. 2023, Feb;30(1), 11-18, doi: 10.1007/s10140-022-02092-8.

Acknowledgement

I would like to express my sincere gratitude to everyone who supported and advised me during the completion of this doctoral dissertation.

First of all, I would like to extend my deepest thanks to my supervisor, Professor Tamotsu Kamishima for his great guidance from the invaluable insights. I always received accurate feedback, which helped me clarify and deepen the direction of my study. It has been a pleasure to work with and learn from him. I sincerely appreciate to Professor Harukazu Tohyama for his guidance on how to improve this study from an orthopedic surgeon's professional perspective. I would like to thank Professor Akihiro Ishizu, Professor Harukazu Tohyama, and Professor Koichi Yokosawa, members of the review committee for this study, for their academic comments and advice.

Then, I am thankful to Dr. Yasuka Kikuchi and Dr. Tasuku Kimura for their assistance in the visual evaluation to determine the ground truth of the lumbar vertebrae conditions in image data used in this study. Also, I am thankful to Ms. Makoto Oda and Mr. Ryosuke Sakano of Hokkaido University Hospital for their cooperation in collecting radiographs. I am very grateful to Dr. Kenneth Sutherland not only for helping me correct my grammatical mistakes and make the article more fluent and natural but also for giving me expert advice on the methodology of deep learning. I would like to thank Mr. Nobuaki Suzuki, former chief radiological technologist, and my colleagues at the workplace. For their understanding of my research activities, I was able to balance work and research.

Finally, I express my sincere gratitude to my family for their understanding, daily generous support, and all the time encourages.

業績リスト

1. 著書

なし

2. 学会誌又は学術雑誌への論文掲載

Ⅰ. 論文発表

・A Deep Learning-Based Model for Classifying Osteoporotic Lumbar Vertebral Fractures on Radiographs: A Retrospective Model Development and Validation Study, Yohei Ono, Nobuaki Suzuki, Ryosuke Sakano, Yasuka Kikuchi, Tasuku Kimura, Kenneth Sutherland, Tamotsu Kamishima, Journal of Imaging, 9(9):187, 2023

Ⅱ. 口頭発表

・深層学習を用いた単純 X 線写真による腰椎圧迫骨折新旧判定への試み，小野陽平，鈴木信昭，坂野稜典，神島保，第 49 回日本放射線技術学会秋季学術大会，熊本，2021 年 10 月

・Classification of Fresh and Old Lumbar Compression Fractures Using Deep Learning Methods on Radiography, Yohei Ono, Nobuaki Suzuki, Ryosuke Sakano, Tamotsu Kamishima, The 5th FHS International Conference, Sapporo, September 2021

・腰椎圧迫骨折の新旧に関する CNN 分類－腰椎自動検出の精度評価と分類能に与える影響，小野陽平，鈴木信昭，坂野稜典，神島保，第 78 回日本放射線技術学会総会学術大会，神奈川，2022 年 4 月

3. 総説・解説

なし

4. 学会賞・学術賞の受賞

なし

5. その他

なし