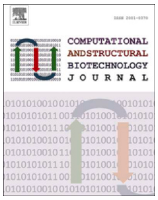
Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj



Research article

# Diagnosis of osteoporotic vertebral compression fractures and fracture level detection using multitask learning with U-Net in lumbar spine lateral radiographs

Seung Min Ryu [a,b,1], Soyoung Lee [b,1], Miso Jang [c], Jung-Min Koh [d], Sung Jin Bae [e], Seong Gyu Jegal [b], Keewon Shin [b,*,2], Namkug Kim [b,c,**,2]

[a] *Department of Orthopedic Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea*
[b] *Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea*
[c] *Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea*
[d] *Division of Endocrinology and Metabolism, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea*
[e] *Department of Health Screening and Promotion Center, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 3 January 2023
Received in revised form 18 June 2023
Accepted 24 June 2023
Available online 27 June 2023

*Keywords:*
Deep Learning
Computer-Assisted Diagnosis
Computer-Assisted Radiographic Image Interpretation
Compression Fractures
Osteoporotic Fractures

ABSTRACT

Recent studies of automatic diagnosis of vertebral compression fractures (VCFs) using deep learning mainly focus on segmentation and vertebral level detection in lumbar spine lateral radiographs (LSLRs). Herein, we developed a model for simultaneous VCF diagnosis and vertebral level detection without using adjacent vertebral bodies. In total, 1102 patients with VCF, 1171 controls were enrolled. The 1865, 208, and 198 LSLRS were divided into training, validation, and test dataset. A ground truth label with a 4-point trapezoidal shape was made based on radiological reports showing normal or VCF at some vertebral level. We applied a modified U-Net architecture, in which decoders were trained to detect VCF and vertebral levels, sharing the same encoder. The multi-task model was significantly better than the single-task model in sensitivity and area under the receiver operating characteristic curve. In the internal dataset, the accuracy, sensitivity, and specificity of fracture detection per patient or vertebral body were 0.929, 0.944, and 0.917 or 0.947, 0.628, and 0.977, respectively. In external validation, those of fracture detection per patient or vertebral body were 0.713, 0.979, and 0.447 or 0.828, 0.936, and 0.820, respectively. The success rates were 96 % and 94 % for vertebral level detection in internal and external validation, respectively. The multi-task-shared encoder was significantly better than the single-task encoder. Furthermore, both fracture and vertebral level detection was good in internal and external validation. Our deep learning model may help radiologists perform real-life medical examinations.

## 1. Introduction

Lumbar spine lateral radiographs (LSLRs), with or without anteroposterior views, are commonly used to diagnose vertebral compression fractures (VCFs) [1]. LSLR is a widely used and cost-effective diagnostic tool for VCF screening [2]. Radiographic criteria for VCFs include at least a 20 % compression in vertebral body height or a 4 mm reduction from baseline height [3]. In postmenopausal women, VCFs are the most prevalent osteoporotic fractures, and osteopenic or osteoporotic bones make it challenging to detect

[*] Correspondence to: Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, 4F, 26, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea.

[**] Correspondence to: Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, 5F, 26, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea.

*E-mail addresses:* kevinkwshin@gmail.com (K. Shin), namkugkim@gmail.com (N. Kim).

[1] These authors contributed equally (co-first authors).

[2] These authors contributed equally (co-corresponding authors).

https://doi.org/10.1016/j.csbj.2023.06.017

subtle VCFs [4]. In addition, radiologists perform the tedious chore of screening for VCFs; therefore, automated diagnosis is needed [5].

Recent advances in deep learning enable automatic diagnosis of a variety of illnesses using two-dimensional radiographs [6,7]. Particularly, studies using deep learning in LSLRs have recently been published. Seo et al. reported dual deep learning models for vertebral body segmentation and vertebral compression measurement using adjacent vertebral bodies [8]. Kim et al. reported automated segmentation and lumbar vertebral level detection models using LSLRs [5]. Kim et al. reported another automated vertebral segmentation-based manual vertebral compression ratio measurement model using LSLRs [9]. Li et al. reported automated vertebral compression fracture detection in LSLRs using you only look once (YOLO) version 3 detection and other classification ensemble models [10]. However, upon analysis, these previous studies mainly focused on segmentation and vertebral level detection in LSLRs [5,8,9]. In addition, previous studies had the disadvantage of making a diagnosis only when clinically meaningful VCFs were discontinuous. In other words, the diagnosis could only be made when the vertebral bodies above and below the fractured vertebral body were normal.

Unlike previous studies, this study first attempted to improve the performance of vertebral body segmentation; second, it attempted to develop a VCF diagnosis model that is not dependent on adjacent bones; and third it attempted to develop a model of vertebral level detection, simultaneously. Finally, we intended to develop a deep learning model that diagnoses the level of VCF to help radiologists perform actual readings.

## 2. Methods

### 2.1. Dataset

This retrospective study was conducted according to the principles of the Declaration of Helsinki and in accordance with the current scientific guidelines. The research protocol was approved by the Institutional Review Board of our institution (S2019–2003–0005), which waived the requirement for informed consent considering the retrospective nature of the study and deidentification of the characteristics of the dataset in accordance with the Health Insurance Portability and Accountability Act privacy rule. A total of 83,005 people with LSLRs from medical check-ups performed between 2011 and 2019 at our institution were reviewed. (Fig. 1 and Supplemental Fig. 1) Of the 83,005 people, 1102 patients with VCF were enrolled, and 1171 controls were selected; therefore, a total of 2273 patients were enrolled.

Normal individuals in their 40 s were intentionally selected for the control group to make deep learning more effective. (Supplemental Fig. 1C) Furthermore, the 2073 LSLRs obtained earlier were selected and randomly split into 1865 and 208 for training and validation. Two hundred consecutive LSLRs obtained afterward were used as the test dataset. In the test set, two LSLRS were excluded due to severely fused vertebrae. (Supplemental Fig. 2) The fractured group had an average of 2.04 VCFs (95 % confidence interval [CI] 1.96–2.11)

We used VinDr-SpineXR images for external validation [11]. In this public data, the training and test sets were separated originally. Then, the label was grouped based on "vertebral collapse" and "no finding". Among them, the number of LSLR images of VCF cases with identifiable ages, sex, and pixel-spacing information was 47. The same number of controls were also selected and matched according to age and sex.

### 2.2. Ground truth labeling

The ground truth label of the internal dataset was made using ITK-SNAP based on radiological reports of normal findings or fractures at some vertebral level [12]. (Supplemental Fig. 3) We used case only when the word "fracture" was reported. The labels were divided into five classes: background, lumbar, thoracic, fracture, and sacrum. Based on the radiologist’s report, a professional labeler with years of experience conducted the entire labeling using a trapezoid shape.

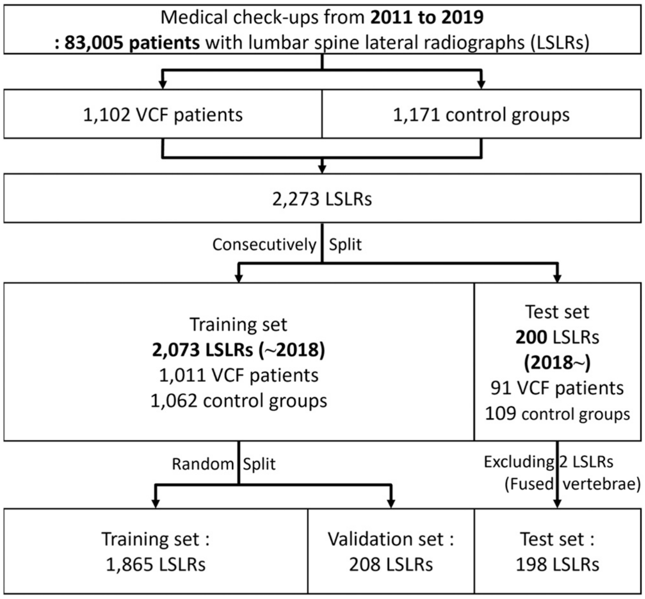


**Fig. 1.** Flowchart summarizing the study design.

### 2.3. Definition of segmentation and evaluation methods

For vertebral body segmentation, performance was evaluated using a Dice score [13,14]. The ground truth labeling for detecting VCF was categorized into four classes: background, VCF, other normal (nonfractured) vertebral bodies, and sacrum. A Dice loss was used as a loss function. Output was evaluated using accuracy, sensitivity, and specificity per patient.

For vertebral level detection, the ground truth was categorized into four classes: background, thoracic spine, lumbar spine (L-spine), and sacrum. The LSLRs of VCF were excluded in the classes. The Dice loss function was also used. The accuracy of vertebral level detection was evaluated for normal examinees and patients with VCF.

### 2.4. Image preprocessing

There were four steps in image preprocessing. First, intensity outlier clipping was applied to 1 % of the upper margins to remove the left/right mark, metal implant, etc.; z-score normalization (the mean of 3028.37 and standard deviation of 1720.31 were calculated for the intensity outlier clipped training set) was then performed. Next, resizing to 1024 × 1024 with zero padding and contrast limited adaptive histogram equalization (CLAHE) were performed [15], and finally, the images were saved as portable network graphic (PNG) files [16] from digital imaging and communications in medicine (DICOM) files [17]. (Supplemental Fig. 4) The image mask was also transformed into the same size as the LSLR. In addition, the mask was modified according to the VCF and level detection tasks as mentioned in Section 2.3. ***Definition of segmentation and evaluation methods***.

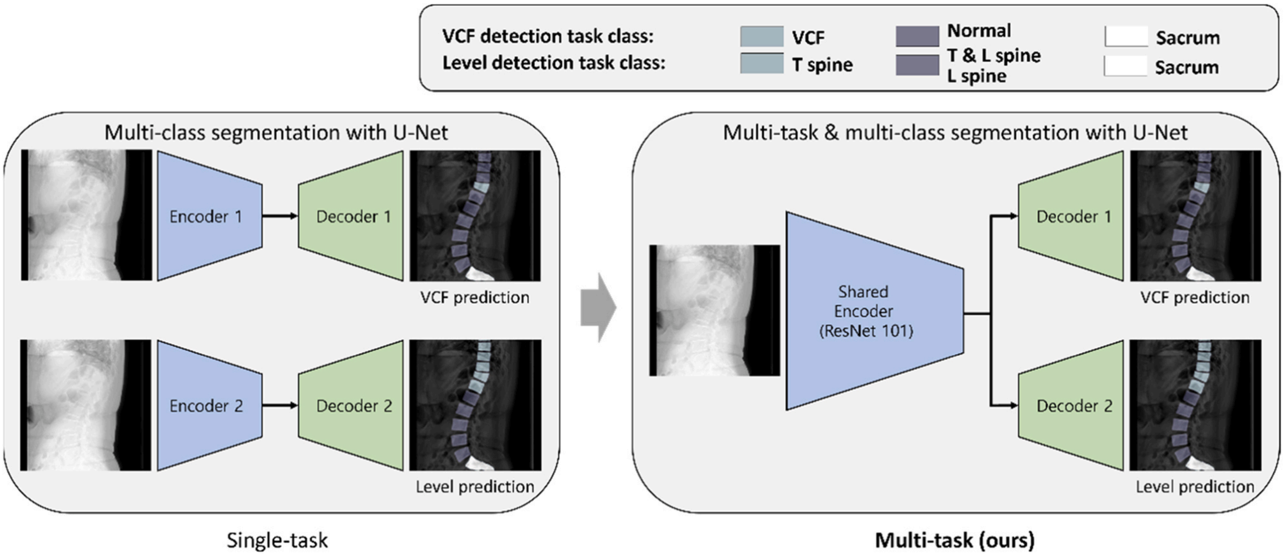


**Fig. 2.** Schematic diagram of the multi-task model (ours) compared with single-task models.

### 2.5. Deep learning architecture and experiment setting

We applied modified the U-Net architecture [18], in which the single encoder is ResNet [19], and the decoders were trained for two tasks [20], including VCF and vertebral level detection (Fig. 2). Because our objective was to obtain the results of two multiclass segmentation tasks using the same radiographs, we designed the model to share the encoder, as shown on the right, to achieve synergy. Each decoder output has four channels (task classes and background), and the class of one pixel is decided by taking the softmax function and taking the largest value channel. In this model, we used Adam optimizer (in PyTorch [21]), the initial learning rate of 1e-4, and ExponentialLR scheduler [22] with gamma of 0.9. We also trained the model for 100 epochs early stopping patience of 5.

ResNet101, ResNet34, ResNet50, and ResNet152 are all ResNet architectures. These ResNet architectures were designed to solve the problem of gradients vanishing in very deep layers, which can inhibit the training process of the network.

These 4 ResNet architectures differ in the number of layers and filters used. Therefore, the numbers, in the end, represent the number of layers, respectively. As a result, more layers tend to have higher representational power and can achieve better performance on tasks with large and complex datasets [23,24]. However, they also require more computational resources and may need to train faster than fewer layers. In this study, we ablated these 4 ResNet architectures and selected the most powerful and efficient layers among these architectures.

### 2.6. Image postprocessing at inference stage

The opening method was used for both tasks to remove small clusters, with a 5-by-5 kernel and 5 iterations based on OpenCV [25]. We then used the criteria to eliminate stains such that each vertebral body had just one class. Finally, we set the number of vertebral bodies in the L-spine to five. Through experimentation, the 30 % stain removal threshold was defined. (Supplemental Figure 5).

### 2.7. Detecting the center of the vertebral body

For the quantitative evaluation of our model, we adopted the Euclidian distance error between the model's inference and the ground-truth centers of the vertebral bodies considering pixel spacing. The error was determined for the cases where the five lumbar vertebrae were successfully detected. Incorrect cases were excluded during calculation.

The four corners of the lumbar vertebral bodies in the test dataset can be determined so that the ground truth mask label can be determined. The center of mass of each mask could be calculated using SciPy [26], and the order from L1 to L5 could be determined depending on the superior to inferior location. The mean square error between the same numbering dots from the ground truth mask and the interference mask was determined as the Euclidean distance by reflecting pixel spacing.

### 2.8. Statistical analysis

The means and 95 % CIs were calculated using Python with Numpy and Pandas libraries. Paired *t*-test and area under the receiver operating characteristic curve (AUROC) analyses were performed using SciPy (version 1.6.3) library [26]. DeLong's method was used for the comparison between AUROCs [27]. The AUROC value was calculated by changing the threshold value of stain removal. Levels of statistical significance that can be mentioned without further explanation are indicated using * ($P$ < 0.05) in the tables.

## 3. Results

### 3.1. Fracture detection per patient

The performances of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) per patient are listed in the test dataset in Table 1. The accuracy, sensitivity, specificity, and F1 scores are also listed in Table 1. Although ResNet152 has more layers than Resnet101, it does not necessarily mean that it is more complex or has more parameters. In fact, Resnet101 has about 44.5 million parameters, while ResNet152 has about 60.2 million parameters [28]. Therefore, we believe that Resnet101 is a more efficient backbone for multi-task learning than ResNet152, as it achieves similar or better performance with fewer parameters. In our model (ResNet101 with multi-task), accuracy, sensitivity, and specificity were 0.929, 0.944, and 0.917, respectively, which were higher than 0.9. Based on the AUROC values, the multi-task model was significantly better than the single-task model.

**Table 1**
Ablation studies for fracture detection per patient (198 cases).

| Models | Performances | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TP | FP | FN | TN | Accuracy | Sensitivity | Specificity | F1 score | DOR | AUROC† |
| Within ResNet101 | | | | | | | | | | |
| **ResNet101 - multi-task (ours)** | 84 | 9 | 5 | 100 | 0.929 | 0.944 | 0.917 | 0.923 | 186.667 | 0.953 |
| ResNet101 - single-task | 77 | 9 | 12 | 100 | 0.894 | 0.865 | 0.917 | 0.880 | 71.296 | 0.906* |
| Within multi-task | | | | | | | | | | |
| **ResNet101 (ours)** | 84 | 9 | 5 | 100 | 0.929 | 0.944 | 0.917 | 0.923 | 186.667 | 0.953 |
| ResNet34 | 82 | 19 | 7 | 90 | 0.869 | 0.921 | 0.826 | 0.863 | 55.489 | 0.908* |
| ResNet50 | 81 | 8 | 8 | 101 | 0.919 | 0.910 | 0.927 | 0.910 | 127.828 | 0.934 |
| ResNet152 | 84 | 9 | 5 | 100 | 0.929 | 0.944 | 0.917 | 0.923 | 186.667 | 0.962 |

Note: TP, True positive; TN, True negative; FP, False positive; FN, False negative; DOR, Diagnostic odds ratio; AUROC, Area under the receiver operating characteristic curve †Statistically significant differences between our model and others in the comparison between AUROCs using DeLong's method are indicted using *($P < 0.05$) in the tables.

**Table 2**
Ablation studies for fracture detection per vertebral body (2127 vertebral bodies).

| Models | Numbers | | Performances | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Vertebral body | Fractured | TP | FP | FN | TN | Accuracy† | Sensitivity† | Specificity† | F1 score† | DOR | AUROC‡ |
| Within ResNet101 | | | | | | | | | | | | |
| **ResNet101 – multi-task (ours)** | 2127 | 183 | 115 | 44 | 68 | 1900 | 0.947 | 0.628 | 0.977 | 0.673 | 73.028 | 0.861 |
| ResNet101 – single-task | 2127 | 183 | 105 | 44 | 78 | 1900 | 0.943 | 0.574* | 0.977 | 0.633 | 58.129 | 0.826 |
| Within multi-task | | | | | | | | | | | | |
| **ResNet101 (ours)** | 2127 | 183 | 115 | 44 | 68 | 1900 | 0.947 | 0.628 | 0.977 | 0.673 | 73.028 | 0.861 |
| ResNet34 | 2127 | 183 | 111 | 48 | 72 | 1896 | 0.944 | 0.607 | 0.975 | 0.649 | 60.896 | 0.855 |
| ResNet50 | 2127 | 183 | 107 | 40 | 76 | 1904 | 0.945 | 0.585 | 0.979 | 0.648 | 67.016 | 0.824 |
| ResNet152 | 2127 | 183 | 115 | 46 | 68 | 1898 | 0.946 | 0.628 | 0.976 | 0.669 | 69.779 | 0.866 |

Note: TP, True positive; TN, True negative; FP, False positive; FN, False negative; DOR, Diagnostic odds ratio; AUROC, Area under the receiver operating characteristic curve †Statistically significant differences between our model and others in the paired T test are shown using *($P < 0.05$) in the tables. ‡Statistically significant differences between our model and others in the comparison between AUROCs using DeLong's method are shown using *($P < 0.05$) in the tables.

### 3.2. Fracture detection per vertebral body

The performances of TP, TN, FP, and FN per vertebral body are listed in Table 2. The accuracy, sensitivity, and specificity were 0.947, 0.628, and 0.977, respectively. One group example is shown in Fig. 3**A**. Based on sensitivity, the multi-task model was significantly better than the single-task model.

### 3.3. Level detection per patient

The success rate was 96 % for 198 test datasets. Incorrect cases were excluded. The distance errors (mm) are shown in Table 3. The error of center point prediction in our model was less than 1 mm for all lumbar vertebrae. One group example is shown in Fig. 3**B**.

### 3.4. Segmentation performance

The pixel based semantic segmentation performance is presented in Supplemental Table 1 and Supplemental Figure 6. The Dice

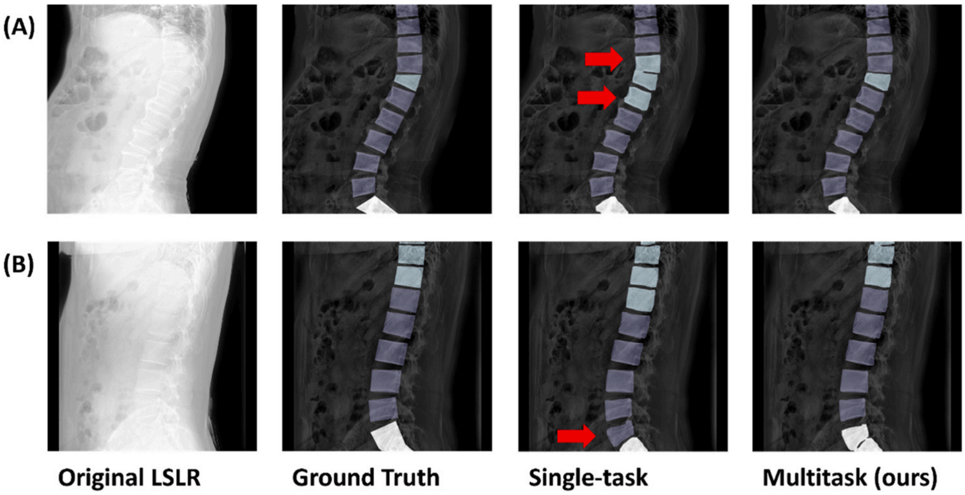


**Fig. 3.** Two cases of **(A)** fracture detection and **(B)** level detection using the models in internal validation. Red arrows indicate the wrong classification. LSLR: lumbar spine lateral radiograph.

**Table 3**
Vertebral level detection results per patient (198 cases).

| Models | Performances | | | Distance errors (mm) in corrected cases | | | | |
|---|---|---|---|---|---|---|---|---|
| | Correct | Incorrect | Accuracy | L1 | L2 | L3 | L4 | L5 |
| Within ResNet101 | | | | | | | | |
| **ResNet101 – multi-task (ours)** | 190 | 8 | 0.960 | 0.86<br>(0.77–0.95) | 0.65<br>(0.59–0.72) | 0.72<br>(0.65–0.79) | 0.75<br>(0.68–0.83) | 0.99<br>(0.89–1.08) |
| ResNet101 – single-task | 184 | 14 | 0.929 | 1.04<br>(0.72–1.35) | 0.89<br>(0.55–1.22) | 0.87<br>(0.46–1.29) | 0.94<br>(0.56–1.31) | 1.19<br>(0.80–1.58) |
| Within multi-task | | | | | | | | |
| **ResNet101 (ours)** | 190 | 8 | 0.960 | 0.86<br>(0.77–0.95) | 0.65<br>(0.59–0.72) | 0.72<br>(0.65–0.79) | 0.75<br>(0.68–0.83) | 0.99<br>(0.89–1.08) |
| ResNet34 | 182 | 16 | 0.919 | 0.83<br>(0.75–0.91) | 0.71<br>(0.64–0.79) | 0.74<br>(0.67–0.81) | 0.83<br>(0.75–0.91) | 1.09<br>(0.99–1.19) |
| ResNet50 | 183 | 15 | 0.924 | 0.86<br>(0.77–0.95) | 0.65<br>(0.59–0.72) | 0.72<br>(0.65–0.79) | 0.75<br>(0.68–0.83) | 0.99<br>(0.89–1.08) |
| ResNet152 | 189 | 9 | 0.955 | 0.91<br>(0.63–1.2) | 0.82<br>(0.50–1.14) | 0.88<br>(0.47–1.29) | 0.92<br>(0.54–1.30) | 1.13<br>(0.72–1.54) |

score of our models were 0.941 and 0.947 in fracture detection decoder and level detection decoder, respectively. The accuracies of our models were 0.989 and 0.991, respectively.

### 3.5. External validation

The fracture detection results per patient are listed on the external dataset in Supplemental Table 2. (Fig. 4**A**) Our model had an accuracy of 0.713, sensitivity of 0.979, and specificity of 0.447. The result of external validation per patient had higher sensitivity and more FP than the result of internal validation.

The fracture detection results per vertebral body are listed in Supplemental Table 3. Our model had an accuracy of 0.828, sensitivity of 0.937, and specificity of 0.820. The external validation result per vertebral body also had higher sensitivity and more FPs than the internal validation result.

The success rate of vertebral level detection was 94 % for 94 external validation datasets. Incorrect cases were also excluded. The distance errors in mm are shown in Supplemental Table 4. One group example is shown in Fig. 4**B**. The center point prediction of our model was also less than 1 mm for all lumbar vertebrae. Similar distance differences were also obtained during internal validation.

The pixel based semantic segmentation performance is presented in Supplemental Table 5. The Dice scores of our models were 0.911 and 0.910 in the fracture and level detection decoders, respectively. The accuracies of our models were both 0.991. The Dice score was approximately 0.03 lower in the external validation than in the internal validation. However, the accuracy was similar.

## 4. Discussion

In this study, we developed a multi-task deep learning model using LSLRs. Our model had favorable outcomes, with 0.9 or higher values in all metrics, including accuracy, sensitivity, specificity, and F1 score per patient. Our model also showed a high specificity of 0.977 when evaluated per vertebral body. In addition, our model showed good accuracy of 96 % in vertebral level detection.

In this study, we proved that multi-task models are superior to single-task models, with values of 0.929 versus 0.894 for accuracy and 0.923 versus 0.880 for F1 scores per patient, respectively, within the same resnet101 backbone. In addition, the values were 0.947 versus 0.943 for accuracy and 0.673 versus 0.633 for F1 score per vertebral body, respectively. By utilizing all information to maximize the performance of a single goal or the general performance of all tasks, multi-task learning seeks to learn many tasks simultaneously. This method generally improves performance and benefits tasks that have similar characteristics [29,30].

All errors from L1 to L5 were less than 1 mm when our model was used to compare the Euclidian distance error of the mass center of the vertebral body. Furthermore, our model showed better results than that of a study by Kim et al., in which the mean Euclidian distance error from L1 to L5 was 4.83–5.37 mm. This study used a

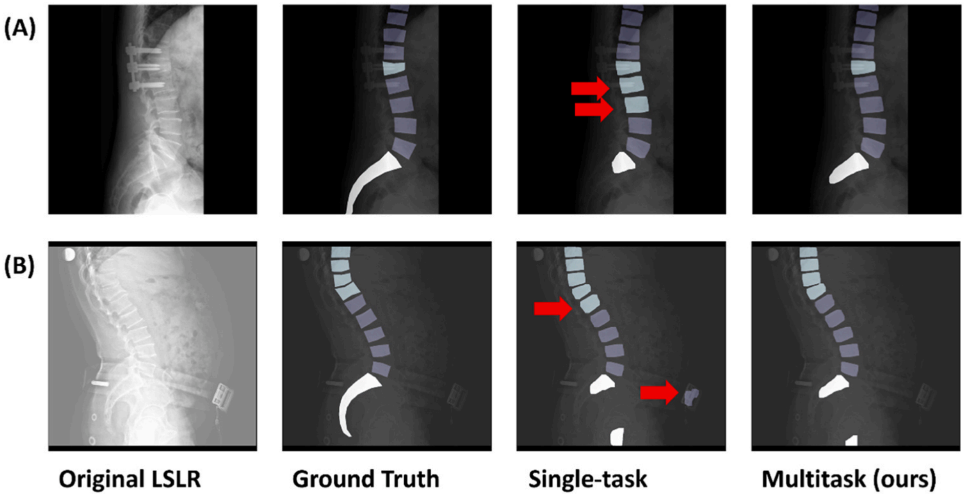


**Fig. 4.** Two cases of **(A)** fracture detection and **(B)** level detection using the models in external validation. Red arrows indicate the wrong classification or segmentation. LSLR: lumbar spine lateral radiograph.

deep learning-based pose estimation method to determine the mass center of the vertebral body [31–33]. The center point estimation using pose estimation method is inherently different from the ground truth center position. Since our model determined the mass center using segmentation and four corners, we assume that our model was slightly better.

In our model, accuracy, sensitivity, specificity, and AUROC per vertebral body were 0.947, 0.628, 0.977, and 0.861, respectively, which are comparable to that in the study by Li et al. that reported 0.93, 0.91,and 0.93, respectively [10]. (Supplemental Table 6) The sensitivity and accuracy were higher in our model than in this model, but the specificity was lower. This might be due to data imbalance. Since the dataset of this study was based on medical check-up data from healthy people who did not know they had a fracture, the distribution of the data is different from that of a study based on only fracture patients, such as Li's study. In addition, our model performed well with accuracy, sensitivity, and specificity of 0.929, 0.944, and 0.917 per patient, respectively. This is probably due to the different perspectives of our model and a professional labeler in multi-level fracture. Higher performance at the patient level would be helpful, as it would give radiologists a chance to take a second look at the X-ray.

The maximum average Dice score of semantic segmentation using our model was 0.947. This was considered to be slightly superior to the 0.923 reported by Seo et al. [8], 0.916 reported by Kim et al. [5], and 0.929 reported by Kim et al. [9]. (Supplemental Table 7) However, the fundamental limit of this segmentation result could not be perfect because the ground truth label was made based on a 4-point square. The vertebral body is not a perfect square and may be distorted by disc degeneration or osteophyte formation [34]; therefore, using a square-based label will not yield perfect results. However, the results of the Dice score were close to 0.95, showing maximum performance given the resources of this study. Nevertheless, we believe that the Dice score was higher than that of other studies because learning was performed using a multi-task model.

The level detection using our model showed 96 % accuracy in 198 test sets, similar to the 96.25 % in 160 test sets reported by Kim et al. [5]. (Supplemental Table 8) We believe that the iliac wing passing between the L4–L5 vertebral bodies could have affected the accuracy, and better accuracy could be obtained if the deep learning model was trained with iliac wing annotations.

Our research had certain limitations. First, the ratio of patients with VCF to the controls was set to approximately 1:1 for effective learning without reflecting the natural prevalence of VCF. In addition, since the ratio of patients in the VCF and control groups was also set to 1:1 for external validation, the real-world scenario was also not reflected. Second, the diagnosis of VCF was not based on CT or MRI, just LSLR. Third, this study does not discriminate between acute and chronic fractures. Third, the fracture detection sensitivity per vertebral body of 0.628 will need to be improved in further studies. Finally, this model may not be technically up to date. However, clinically, it is the first deep learning model to go beyond detecting fractures in 2D X-rays and infer what level they are at.

## 5. Conclusions

In this study, we developed a deep learning model for detecting VCF and vertebral levels using LSLRs. In the process, it was proven that the multi-task-shared encoder was superior to the single-task encoder. In addition, the results of fracture level and vertebral level detection were favorable for internal and external validations. In conclusion, our deep learning model would help radiologists perform real-life medical examinations.

## Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00214256). This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT). (Project No. 2021000393 , Anonymization for Image and Multi-media data that removes Personally Identifiable Information from the data while preserving their AI utility).

## CRediT authorship contribution statement

**Seung Min Ryu (SMR):** Conceptualization, Methodology, Writing - Original Draft, Funding acquisition. **Soyoung Lee (SL):** Conceptualization, Methodology, Software Programming, software development, Resources, Data Curation, Writing - Review & Editing, Visualization. **Miso Jang (MJ):** Conceptualization, Investigation, Resources, Supervision. **Jung-Min Koh (JK):** Conceptualization, Investigation, Resources, Supervision. **Sung Jin Bae (SJB):** Conceptualization, Investigation, Resources, Supervision. **Seong Gyu Jegal (SGJ):** Conceptualization, Investigation, Resources, Supervision. **Keewon Shin (KS):** Conceptualization, Methodology, Software Programming, software development, Resources, Data Curation, Writing - Review & Editing, Visualization. **Namkug Kim (NK):** Conceptualization, Investigation, Resources, Supervision, Project administration.

## Data availability

Our proposed models and all the source codes are publicly available at: (https://github.com/SoyoungLee8/TLspine_FxLevel_Segmentation/tree/main).

## Declaration of Competing Interest

None Declared.

## Acknowledgement

We would like to thank Editage (www.editage.co.kr) for English language editing.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.06.017.

## References

[1] Kiel D. Assessing vertebral fractures. National Osteoporosis Foundation Working Group on Vertebral Fractures. J Bone Min Res 1995;10:518–23. https://doi.org/10.1002/jbmr.5650100403

[2] Wong CC, McGirt MJ. Vertebral compression fractures: a review of current management and multimodal therapy. J Multidiscip Health 2013;6:205–14. https://doi.org/10.2147/JMDH.S31659

[3] Prather H, Hunt D, Watson JO, Gilula LA. Conservative care for patients with osteoporotic vertebral compression fractures. Phys Med Rehabil Clin N Am 2007;18:577–91. https://doi.org/10.1016/j.pmr.2007.05.008. xi.

[4] Kondo KL. Osteoporotic vertebral compression fractures and vertebral augmentation. Semin Interv Radio 2008;25:413–24. https://doi.org/10.1055/s-0028-1103000

[5] 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 Prog Biomed 2021;200:105833. https://doi.org/10.1016/j.cmpb.2020.105833

[6] Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, et al. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287:313–22. https://doi.org/10.1148/radiol.2017170236

[7] Ryu SM, Shin K, Shin SW, Lee S, Kim N. Enhancement of evaluating flatfoot on a weight-bearing lateral radiograph of the foot with U-Net based semantic segmentation on the long axis of tarsal and metatarsal bones in an active learning manner. Comput Biol Med 2022;145:105400. https://doi.org/10.1016/j.compbiomed.2022.105400

[8] Seo JW, Lim SH, Jeong JG, Kim YJ, Kim KG, et al. A deep learning algorithm for automated measurement of vertebral body compression from X-ray images. Sci Rep 2021;11:13732. https://doi.org/10.1038/s41598-021-93017-x

[9] 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;34:853–61. https://doi.org/10.1007/s10278-021-00471-0

[10] Li YC, Chen HH, Horng-Shing Lu H, Hondar Wu HT, Chang MC, et al. Can a deep-learning model for the automated detection of vertebral fractures approach the performance level of human subspecialists. Clin Orthop Relat Res 2021;479:1598–612. https://doi.org/10.1097/CORR.0000000000001685

[11] Pham H.H., Trung H.N., Nguyen H.Q. VinDr-SpineXR: A large annotated medical image dataset for spinal lesions detection and classification from radiographs. In press.

[12] Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006;31:1116–28. https://doi.org/10.1016/j.neuroimage.2006.01.015

[13] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.

[14] Sudre CH, Li WQ, Vercauteren T, Ourselin S, Cardoso MJ. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support 2017;10553:240–8. https://doi.org/10.1007/978-3-319-67558-9_28

[15] Reza A.M. (2004) Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for real-time image enhancement. J Vis Signal Process Syst Signal Image Vid Technol 38:35–44. 10.1023/B:Vlsi.0000028532.53893.82.

[16] Crocker L.D. (1995) Png - the Portable Network Graphic Format. Dr Dobbs Journal 20:36-&.

[17] Mildenberger P, Eichelberg M, Martin E. Introduction to the DICOM standard. Eur Radiol 2002;12:920–7. https://doi.org/10.1007/s003300101100

[18] Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, et al. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognit 2020;106:107404.

[19] He KM, Zhang XY, Ren SQ, Sun J. Deep Residual Learning for Image Recognition. 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr). In press,; 2016. p. 770–8. https://doi.org/10.1109/Cvpr.2016.90

[20] Mehta S, Mercan E, Bartlett J, Weaver D, Elmore JG, et al. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. Medical Image Computing and Computer Assisted Intervention - Miccai 2018 2018;11071:893–901. https://doi.org/10.1007/978-3-030-00934-2_99

[21] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, et al. Pytorch: An imperative style, high-performance deep learning library 2019;32:8026–37.

[22] Li Z., Arora S.Japa (2019) An exponential learning rate schedule for deep learning. In press.

[23] Rawat W, Wang ZH. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput 2017;29:2352–449. https://doi.org/10.1162/neco_a_00990

[24] Simonyan K., Zisserman A Japa (2014) Very deep convolutional networks for large-scale image recognition. In press.

[25] Culjak I, Abram D, Pribanic T, Dzapo H, Cifrek M. A brief introduction to OpenCV. 2012 proceedings of the 35th international convention MIPRO. In press; 2012. p. 1725–30.

[26] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2

[27] Delong ER, Delong DM, Clarkepearson DI. Comparing the areas under 2 or more correlated receiver operating characteristic curves - a nonparametric approach. Biometrics 1988;44:837–45. https://doi.org/10.2307/2531595

[28] Zagoruyko S., Komodakis NJapa (2016) Wide residual networks. In press.

[29] Kyung S, Shin K, Jeong H, Kim KD, Park J, et al. Improved performance and robustness of multi-task representation learning with consistency loss between pretexts for intracranial hemorrhage identification in head CT. Med Image Anal 2022;81:102489. https://doi.org/10.1016/j.media.2022.102489

[30] Vandenhende S, Georgoulis S, Van Gansbeke W, Proesmans M, Dai D, et al. Multi-task learning for dense prediction tasks: a survey. IEEE Trans Pattern Anal Mach Intell 2022;44:3614–33. https://doi.org/10.1109/TPAMI.2021.3054719

[31] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. 30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017). In press; 2017. p. 1302–10. https://doi.org/10.1109/Cvpr.2017.143

[32] Toshev A, Szegedy C. DeepPose: Human Pose Estimation via Deep Neural Networks. 2014 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr). In press; 2014. p. 1653–60. https://doi.org/10.1109/Cvpr.2014.214

[33] Wei SE, Ramakrishna V, Kanade T, Sheikh Y. Convolutional Pose Machines. 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr). In press; 2016. p. 4724–32. https://doi.org/10.1109/Cvpr.2016.511

[34] Heggeness MH, Doherty BJ. Morphologic study of lumbar vertebral osteophytes. South Med J 1998;91:187–9. https://doi.org/10.1097/00007611-199802000-00012