Original Article

Deep learning-driven diagnosis of multi-type vertebra diseases based on computed tomography images

Yongjie Wang[1#], Feng Su[2,3#], Qian Lu[1], Wenkai Zhang[1], Tao Liu[1], Yining Tao[2], Shuai Fu[2], Libin Cui[1], Shi-Bao Lu[4], Xueming Chen[1], Zhenyun Shi[2]

[1]Department of Orthopedics, Beijing Luhe Hospital, Capital Medical University, Beijing, China; [2]School of Mechanical Engineering and Automation, Beihang University, Beijing, China; [3]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China; [4]Department of Orthopedics, Xuanwu Hospital, Capital Medical University, Beijing, China

Contributions: (I) Conception and design: Y Wang, F Su, X Chen, Z Shi; (II) Administrative support: None; (III) Provision of study materials or patients: Y Wang, SB Lu, X Chen; (IV) Collection and assembly of data: Y Wang, Q Lu, W Zhang, T Liu, Y Tao, S Fu, L Cui, SB Lu, X Chen; (V) Data analysis and interpretation: Y Wang, F Su, Z Shi; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Zhenyun Shi, PhD. School of Mechanical Engineering and Automation, Beihang University, Xueyuan Road No. 37, Haidian District, Beijing 100191, China. Email: zyshi@buaa.edu.cn; Xueming Chen, PhD. Department of Orthopedics, Beijing Luhe Hospital, Capital Medical University, Xinhuan South Road No. 82, Tongzhou District, Beijing 101149, China. Email: xuemingchen@sina.com.

Background: Osteoporotic vertebral compression fractures (OVCFs) are the most common type of fragility fracture. Distinguishing between OVCFs and other types of vertebra diseases, such as old fractures (OFs), Schmorl's node (SN), Kummell's disease (KD), and previous surgery (PS), is critical for subsequent surgery and treatment. Combining with advanced deep learning (DL) technologies, this study plans to develop a DL-driven diagnostic system for diagnosing multi-type vertebra diseases.

Methods: We established a large-scale dataset based on the computed tomography (CT) images of 1,051 patients with OVCFs from Luhe Hospital and used data of 46 patients from Xuanwu Hospital as alternative hospital validation dataset. Each patient underwent one examination. The dataset contained 11,417 CT slices and 19,718 manually annotated vertebrae with diseases. A two-stage DL-based system was developed to diagnose five vertebra diseases. The proposed system consisted of a vertebra detection module (VDModule) and a vertebra classification module (VCModule).

Results: The training and testing dataset for the VDModule consisted of 9,135 and 3,212 vertebrae, respectively. The VDModule using the ResNet18-based Faster region-based convolutional neural network (R-CNN) model achieved an area under the curve (AUC), false-positive (FP) rate, and false-negative (FN) rate of 0.982, 1.52%, and 1.33%, respectively, in the testing dataset. The training dataset for VCModule consisted of 14,584 and 47,604 diseased and normal vertebrae, respectively. The testing dataset consisted of 4,489 and 15,122 diseased and normal vertebrae, respectively. The ResNet50-based VCModule achieved an average sensitivity and specificity of 0.919 and 0.995, respectively, in diagnosing four kinds of vertebra diseases except for SN in the testing dataset. In the alternative hospital validation dataset, the ResNet50-based VCModule achieved an average sensitivity and specificity of 0.891 and 0.989, respectively, in diagnosing four kinds of vertebra diseases except for SN.

Conclusions: Our proposed DL system can accurately diagnose four vertebra diseases and has strong potential to facilitate the accurate and rapid diagnosis of vertebral diseases.

Keywords: Osteoporotic vertebral compression fracture (OVCF); deep learning (DL); old fracture (OF); Kummell's disease (KD)

Submitted May 18, 2023. Accepted for publication Nov 24, 2023. Published online Jan 02, 2024.
doi: 10.21037/qims-23-685
View this article at: https://dx.doi.org/10.21037/qims-23-685

Introduction

Osteoporosis usually leads to fragility fractures, especially in the vertebrae, hip, distal radius, proximal humerus, and pelvis. Osteoporotic vertebral compression fractures (OVCFs) are the most common fragility fractures and would lead to loss of fractured vertebral height, intractable back pain, decreased cardiopulmonary function, and gastrointestinal dysfunction (1). OVCFs and their accompanying pain may lead to prolonged bed rest, reduced activity, and further loss of bone mass (2). OVCFs also present a major and growing public concern worldwide (3,4). The 1-year mortality rate for patients with OCVFs is higher than that of the general population (5), and the 4-year survival rate is only 50% (6). Moreover, OVCFs require huge investments in medical treatment and nursing, thereby imposing a heavy burden on individuals and society (7).

The onset of OVCFs is insidious, and only 25% of OVCFs have obvious causes, such as resulting from a fall (8). This leads to a much lower consultation rate among patients with OVCFs than among those with other osteoporotic fractures (9). Although the consultation rate for OVCFs is increasing year by year, underdiagnosis still occurs (9). A multicenter study conducted across different countries showed a false-negative (FN) rate of OVCFs to as high as 34% (10). This combination of low consultation rate and high FN rate meant that only 28.8% of patients with OVCFs started anti-osteoporosis therapy within 1 year after the occurrence of the OVCFs (11). Timely and appropriate treatment for OVCFs can effectively reduce the refracture risk of other vertebrae and the hip (12).

In clinical practice, several medical imaging techniques, including X-rays, magnetic resonance imaging (MRI), and computed tomography (CT), play an important role in diagnosing OVCFs. In radiological images, OVCFs exhibit obvious loss of height in the anterior, middle, or posterior dimension of the vertebral body. X-rays are the fastest and easiest method to implement and are a primary OVCF screening method. However, the sensitivity and specificity of X-rays are relatively low (10,13). MRI can display the anatomy of the spine and is the gold standard in OVCF diagnosis (14,15); however, MRI is expensive and time-consuming, and cannot achieve a timely diagnosis (16). MRI imaging also has some contraindications, such as in patients with pacemaker devices. CT images can also display vertebral fractures well, with higher accuracy than X-rays, and are easier to obtain than MRI (17). Non-enhanced spine CT is usually considered a standard test for fast exclusion or closer assessment of suspected or known vertebral fractures (18). However, occult fractures without significant vertebral compression have only subtle imaging features on CT (19) and are easily overlooked, leading to FN diagnoses (20).

In recent years, deep learning (DL)-assisted diagnosis has emerged in the medical field and has made significant contributions to medical imaging analysis (21-25). Several DL-assisted OVCF diagnostic systems based on CT images have been reported. For example, Tomita *et al.* developed a coupled DL system to analyze whole-spine non-segmented sagittal CT images from 713 patients with thoracic and lumbar OVCFs and 719 individuals without OVCFs (26). The proposed DL system used a convolutional neural network (CNN) to extract image features and a long short-term memory (LSTM) network to aggregate the features and make the final diagnosis for the full CT scan. The system achieved a sensitivity, specificity, and area under the curve (AUC) of 0.85, 0.96, and 0.91, respectively, in 129 hold-out CT scans. Kolanu *et al.* reported a computer-aided diagnosis (CAD) system developed from 1,696 chest or abdominal CT scans and showed an OVCF diagnostic specificity and sensitivity of 0.92 and 0.54, respectively (27). However, these reported studies only classified the presence or absence of OVCFs throughout the CT scan and did not pinpoint the specific locations of the injured vertebrae.

Many different types of vertebra diseases are encountered in clinical practice. Kummell's disease (KD), also called avascular necrosis of a vertebral body or a sign of avascular necrosis of a vertebral body, is a specific type of OVCF (28). KD usually has collapsed vertebrae and intravertebral vacuum cleft and fluid. Due to obvious intravertebral instability, patients with KD are prone to progress to delayed vertebral body collapse, resulting in significant kyphosis, nerve compression, and delayed neurological dysfunction (29). Conservative treatment is often unsatisfactory in most patients with KD, and more aggressive surgical options are usually required (30). Therefore, early diagnosis and early surgical intervention are important for these patients. Schmorl's node (SN) is defined as herniation of the discs into the vertebral body through the endplate (31), and discrete indentations of the endplates are related to degenerative disc disease. Certain types of SNs or the occurrence of SNs in combination with OVCFs may produce low back pain symptoms similar to OVCFs (32), and medium-sized or large SNs may be misinterpreted as endplate fracture (33). Percutaneous vertebroplasty (PVP) surgery is an effective treatment for vertebral fractures. However, previous surgery (PS) may affect the mechanical characteristics of the spine and have a potential impact on recurrent fractures. PS has a significant highlighting signal in the CT image at the site of cement injection. Therefore, distinguishing among various vertebra diseases is clinically important. Prior literature reported that DL methods can diagnose OVCFs from medical images (26,27). Hence, we hypothesized that CT images contain rich features and well-trained DL models can recognize key features for the diagnosis of multi-types vertebral diseases.

Our aim in the present study was to develop a DL-assisted diagnostic system based on CT images that could achieve a single vertebra-level diagnosis. We further improved the practicability of the system by considering several vertebra diseases, including OVCF, old fractures (OFs), SN, KD, and PS. The proposed system has great potential for improving the efficiency and reliability of the vertebral fracture diagnostic process. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-685/rc).

Methods

Patient cohorts

This retrospective study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was performed in accordance with relevant guidelines and regulations, and approved by the Institutional Review Board of Beijing Luhe Hospital, Capital Medical University (No. 2023-LHKY-022-02). Written informed consent was taken from all individual participants. We collected patients from two institutes: Beijing Luhe Hospital and Xuanwu Hospital. All patients were diagnosed with OVCFs and underwent surgery. All patients underwent X-ray, CT, and MRI examinations. X-rays were used to confirm the location of fractures and locate them during surgery. The aim of this study was to achieve high accuracy in diagnosing vertebral diseases based on CT scans. MRI examinations were used to ensure the reliability of disease diagnosis during the construction of a multi-type disease dataset. Each patient's vertebral fracture diagnosis was made by three senior spine surgeons according to the patient's imaging and clinical history. Each patient's sagittal CT images spanned multiple thoracic and lumbar vertebrae. The acquired sagittal CT images covered the vertebrae from T1 to L5 (Table S1). Details about imaging parameters are shown in supplementary materials (Appendix 1).

The patient cohort from Luhe Hospital consisted of 1,198 patients diagnosed with OVCFs from 1 January 2015 to 31 December 2020. Inclusion criteria were as follows: (I) age ≥55 years; (II) both male and female were considered; (III) patients had undergone X-ray, CT, and MRI examinations; (IV) thoracic or/and lumbar vertebrae were considered. Exclusion criteria were as follows: (I) vertebral fracture due to high violence injury, such as traffic injury or fall from height injury; (II) incomplete imaging data and abnormal image quality; (III) pathological fracture due to primary malignant tumor or metastatic tumor. We analyzed a total of n=1,051 patients (excluding 112 patients with incomplete image data, 21 patients with poor image quality, eight patients with multiple myeloma, and six patients with metastatic tumors).

The training cohort contained 8,548 CT slices from 819 patients, and the testing cohort contained 2,456 CT slices from 232 patients (*Table 1*). The training and testing cohorts showed no significant differences in age, gender distribution, or dual-energy X-ray absorptiometry (DEXA) T-scores. We also collected an alternative Hospital validation cohort from Xuanwu Hospital, containing 467 CT slices of 46 patients from the period of 1 January 2015 to 31 December 2020.

Annotation of the datasets

Three spine surgeons who have been specializing in spine surgery for more than 15 years made vertebra diagnoses according to the patient's spinal CT images and clinical history. For the vertebra detection task, all vertebrae in the CT images were annotated by bounding boxes (Bboxes). All annotations were assigned the same label of "vertebra". For the vertebra classification task, only the vertebrae with injuries were annotated by Bboxes. Five kinds of injured vertebrae were considered: OVCF, OF, SN, KD, and PS. Surgeons annotated each CT slice independently, followed

Table 1 Demographic data of patients

Characteristics	Luhe Hospital (training cohort)	Luhe Hospital (testing cohort)	Xuanwu Hospital (alternative hospital validation cohort)
Patient count	819	232	46
Age (year)	73±8	72±8	76±9
Sex, n (%)			
Male	192 (23.4)	52 (22.4)	13 (28.3)
Female	627 (76.6)	180 (77.6)	33 (71.7)
DEXA T-score	−2.9±1.3	−2.9±1.4	−2.9±0.5
Number of slices	8,548	2,456	467

Data were presented as mean ± SD or n (%). The training cohort, testing cohort, and alternative hospital validation cohort showed no significant differences in gender distribution (Pearson χ^2 test, χ^2=0.732, P=0.693). These datasets also showed no significant differences in DEXA T-score (ANOVA Bartlett's test with Tukey's multiple comparisons; P=0.368). The patients used for alternative hospital validation were older than the patients in the training and testing cohorts from Luhe Hospital (ANOVA Bartlett's test with Tukey's multiple comparisons; P=0.004; Train *vs.* Test P=0.093; Train *vs.* Validation, P=0.003; Test *vs.* Validation, P=0.014). DEXA, dual-energy X-ray absorptiometry; SD, standard deviation; ANOVA, analysis of variance.

by consulted co-annotation for inconsistent initial diagnoses (Figures S1,S2).

Development of the DL-based vertebra diagnostic system

We overcame the limitations in the traditional vertebra disease diagnosis workflow by developing an intelligent DL-based vertebra diagnostic system (*Figure 1A-1E*). The DL workflow of the system consisted of three modules (*Figure 1A*): a vertebra detection module (VDModule), a vertebra extraction module (VEModule), and a vertebra classification module (VCModule). For an input CT slice, the VDModule first detected all vertebrae and calculated the exact Bbox of all vertebrae (*Figure 1B*). The VEModule then extracted multiple image patches according to the determined Bbox using several combined image-processing operations (*Figure 1C*). Finally, the extracted vertebra image patches were passed into the VCModule to achieve the diagnosis (*Figure 1D*). Five vertebra diseases (OVCF, OF, SN, KD, and PS) were considered in this study. Notably, these vertebra diseases are not mutually exclusive, and several may occur in the same vertebra. We accounted for this disease co-occurrence problem and adopted a multi-output DL model in the VCModule (*Figure 1D*). The system was developed and validated using the datasets from the two hospitals (*Figure 1E*).

Development of the VDModule

We developed a VDModule based on Faster region-based CNN (R-CNN) architecture. The Faster R-CNN models were trained to automatically determine the Bboxes of all vertebrae in each CT slice. A pretrained ResNet18 model and MobileNet v2 model were used as the backbone of the Faster R-CNN model. We adopted the input image size of 1,024×1,024 to develop the VDModule. The original sagittal CT slices were usually not square (image height 743±111, image width 619±122). We normalized the image size and avoid image deformation by adopting center-cropping and zero-padding techniques to modify all sagittal CT images to a 1,024×1,024 size. We enlarged the image and Bbox datasets eight-fold using offline data augmentation (Figure S3) (3); this augmentation strategy did not decrease the reliability of ground truth and was proved efficient for detecting cells from cytopathology images (34). Other parameter settings are shown in supplementary material (Appendix 1).

Evaluation of the VDModule

We adopted three different measurements to evaluate the performance of the VDModule. First, we calculated the precision-recall curve and the corresponding mean average precision (mAP) value for each Faster R-CNN model. Second, we set a threshold of 0.75 for the probability of vertebra detection results and compared the number of detected and annotated vertebrae for each image. Third, we compared the spatial relationship between the Bboxes of detected and annotated vertebrae. We calculated the intersection over minimum (IoM) score between two Bboxes

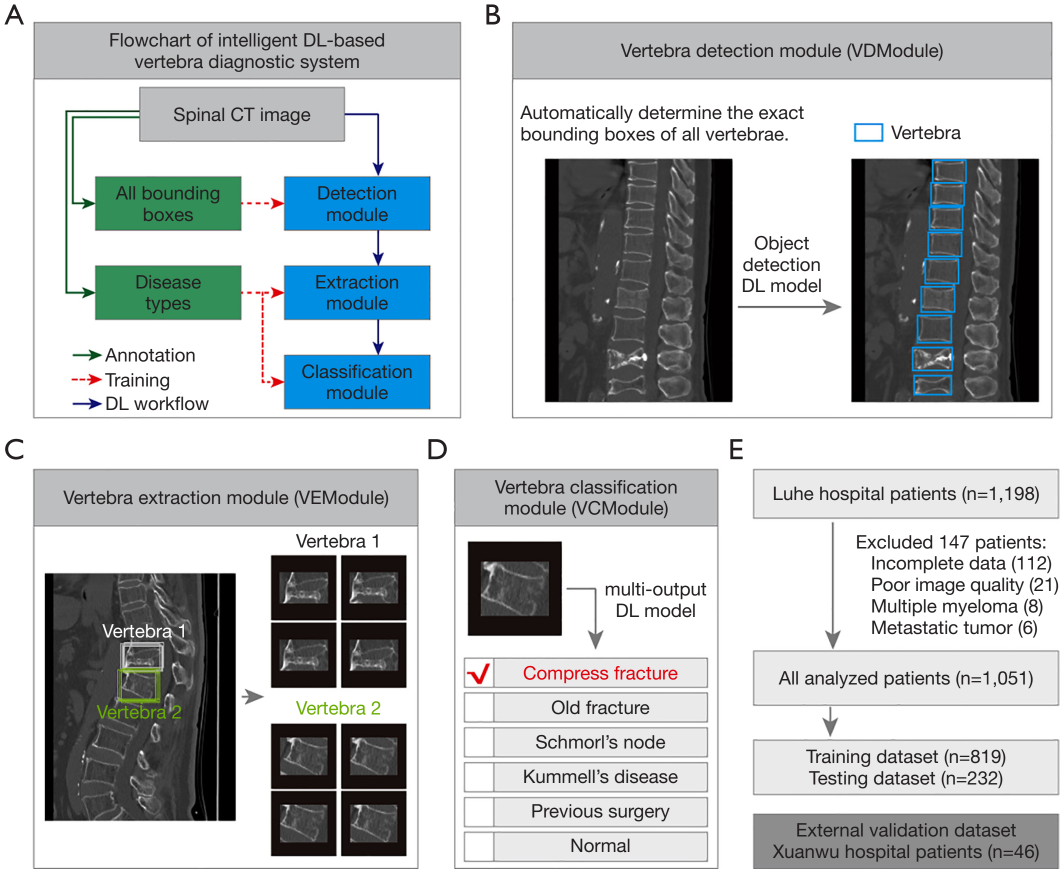

Figure 1 Architecture of intelligent DL-based vertebra diagnostic system. (A) Flowchart for the development of a DL-based vertebra diagnostic system. Expert annotation included two types: labeling all vertebrae by Bboxes for training vertebra detection models and labeling vertebrae with diseases for training vertebra classification models. The DL workflow consists of three modules: a VDModule, a VEModule, and a VCModule. (B) The function of the VDModule. The VDModule processes the clinical CT image and automatically determines the exact Bbox of all vertebrae. Left, clinical CT image. Right, detected vertebrae. The VDModule is developed based on object-detection DL models. (C) The function of VEModule. The VEModule obtains multiple samples from one Bbox of the target vertebra by implementing scaling and random translation, thereby realizing data augmentation. The eight cropped vertebra image patches shown as gray and green Bboxes correspond to vertebrae 1 and 2, respectively. (D) The function of the VCModule. The VCModule is developed based on a multi-output DL model and implements the diagnosis of the input vertebra patch. The vertebra types recognized by VCModule include OVCF, OFs, SN, KD, PS, and normal. (E) Two patient cohorts were used in this study. The training and testing datasets were from Luhe Hospital and the alternative hospital validation dataset was from Xuanwu Hospital. DL, deep learning; CT, computed tomography; VDModule, vertebra detection module; VEModule, vertebra extraction module; VCModule, vertebra classification module; Bboxes, bounding boxes; OVCF, osteoporotic vertebral compression fracture; OF, old fracture; SN, Schmorl's node; KD, Kummell's disease; PS, previous surgery.

as the area of the intersection between two Bboxes divided by the minimum area of the Bboxes. We named a detected Bbox and an annotated Bbox as a hit if their IoM was larger than 0.5. The FN error scenario referred to annotated but not detected vertebrae. The false-positive (FP) error scenario referred to incorrectly detected vertebrae.

Development of the VCModule

We considered six vertebra categories in this study, including five disease categories (OVCF, OF, SN, KD, and PS) and one normal category. The five disease categories were not mutually independent. Hence, we developed a VCModule based on a multi-output DL model. We

Table 2 Dataset of image patches for all classes

Dataset	Compression fracture	Old fracture	Schmorl's node	Kummell's disease	Previous surgery	Normal
Training dataset	8,046 (12.9)	3,257 (5.2)	824 (1.3)	478 (0.8)	1,979 (3.2)	47,604 (76.5)
Testing dataset	2,536 (12.9)	893 (4.6)	154 (0.8)	204 (1.0)	702 (3.6)	15,122 (77.1)
Alternative hospital dataset	398 (13.0)	66 (2.1)	23 (0.7)	56 (1.8)	102 (3.3)	2,425 (79.0)

Data were presented as samples, n (%).

developed the VCModule in two steps. We first constructed the vertebra patch dataset for the VCModule; we used the VDModule to detect all vertebrae, and we eliminated the detected vertebrae that largely overlapped the diseased vertebrae to obtain the normal vertebrae. We then developed vertebra classification models. A pretrained ResNet50 served as the backbone of the DL model using transfer learning techniques. We used random over-sampling and random under-sampling strategies to solve the class imbalance problem, and used data augmentation techniques to solve the overfitting problem. These re-sampling strategies were applied only to the training dataset, and not to the testing dataset and alternative hospital validation dataset. In the original training dataset, the samples for different vertebra diseases were highly imbalanced (counts of vertebra patches: OVCF 8,046; OF 3,257; SN 824; KD 478, PS 1,979; normal 47,604; *Table 2*). We used the count of the OVCF samples, which was the disease category with the most samples, as the reference for the re-sampling operation. The normal vertebra class was the majority class, and we randomly chosen a part of the samples for training using under-sampling. The number of chosen normal vertebrae was twice the OVCF count. The OF, SN, KD, and PS categories were the minority classes, and we randomly duplicated the samples to half the OVCF counts using over-sampling. After re-sampling, we established a more balanced training dataset for the various vertebra categories (counts of vertebra patches: OVCF 8,046; OF 4,023; SN 4,023; KD 4,023; PS 4,023; normal 16,092).

Other parameter settings are shown in supplementary material (Appendix 1).

Evaluation of the VCModule

We adopted three kinds of measurements to evaluate the performance of the VCModule. First, we calculated the one-*vs.*-all confusion matrix of the target and other labels for each vertebra category. Second, we calculated the one-*vs.*-all ROC curve. Third, we calculated several criteria to evaluate the diagnostic performance. For each disease category, true positive (TP) and false negative (FN) means accurately and incorrectly classified diseased vertebrae, respectively, while true negative (TN) and false positive (FP) means accurately and incorrectly classified normal vertebrae, respectively. For each vertebra category, we calculated five quantitative measurements: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. The PPV and sensitivity are also called precision and recall, respectively. The F1 score is the harmonic mean of precision and recall and provides a balanced evaluation of the model's performance.

$$Sensitivity = \frac{TP}{TP + FN} \quad [1]$$

$$Specificity = \frac{TN}{TN + FP} \quad [2]$$

$$PPV = \frac{TP}{TP + FP} \quad [3]$$

$$NPV = \frac{TN}{TN + FN} \quad [4]$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad [5]$$

Statistical analysis

Data were analyzed using GraphPad 7.0 and IBM SPSS Statistics 26 softwares. Data were presented as mean ± standard deviation (SD). Frequency distribution was tested by Pearson Chi-square (χ^2) test. Statistical analysis was performed using one-way analysis of variance (ANOVA), with Bartlett's test and Tukey's multiple comparisons.

Statistical significance was set at P<0.05.

Results

Performance of the VDModule in vertebra detection

We achieved automatic vertebra diagnosis from the original images by first developing a VDModule based on the Faster R-CNN model to detect all the vertebrae in the input image (*Figure 2A*). The training dataset for the VDModule consisted of 1,082 images with 9,135 vertebrae, and the testing dataset consisted of 405 images with 3,212 vertebrae. The vertebrae-level ratio of training and testing datasets for VDModule was approximately 3:1. Both the ResNet18 and MobileNet v2-based models were well trained, with increased region proposal accuracy and decreased loss as the iteration increased (*Figure 2B*). The ResNet18-based faster R-CNN model showed significantly better performance than the MobileNet v2-based faster R-CNN model in vertebra detection task (comparison of the precision-recall curve, Wilcoxon signed rank test, P<0.001; mAP; ResNet18, 0.982; MobileNet v2, 0.941; *Figure 2C*).

We also quantitatively evaluated the relationship between the annotations and the ResNet18-based model detection results. The numbers of manually annotated and model-detected vertebrae in each CT slice were significantly linearly correlated in the testing datasets (*Figure 2D*). We also divided the annotated and model-detected vertebrae into three types (hit, FN, and FP) to assess the reliability of the ResNet18-based model (*Figure 2E*). The developed ResNet18-based model showed high performance with the testing datasets, yielding FP and FN rates of 1.52% and 1.33%, respectively (*Figure 2F*). Our experimental results showed that the proposed VDModule achieved accurate and reliable vertebra detection.

Performance of the VCModule in vertebra classification

In the single vertebra level dataset, the training dataset consisted of 14,584 manually annotated diseased vertebrae and 47,604 semi-automatically detected normal vertebrae (*Figure 3A,3B*), and the testing dataset consisted of 4,489 manually annotated diseased vertebrae and 15,122 semi-automatically detected normal vertebrae (*Table 2*). The vertebrae-level ratio of training and testing datasets for VCModule was approximately 3:1.

We used the established training and testing datasets to develop a multi-output classification model based on ResNet-50 to classify each vertebra into six classes independently: OVCF, OF, SN, KD, PS, and normal (*Figure 3C-3F*). The VCModule exhibited high performance for the OVCF, OF, KD, and PS classes in both the training and testing datasets (training dataset: sensitivity 0.945±0.052, specificity 0.997±0.004, PPV 0.964±0.026, NPV 0.998±0.003, F1 score 0.954±0.033, *Table 3* and Figure S4; testing dataset: sensitivity 0.919±0.082, specificity 0.995±0.006, PPV 0.930±0.037, NPV 0.996±0.004, F1 score 0.924±0.056, *Table 3* and *Figure 3E*). The F1 scores for the training and testing datasets showed no significant differences (*t*-test, P=0.390). However, the VCModule exhibited relatively poor performance for SN diagnosis (training dataset: sensitivity 0.713, specificity 0.998, PPV 0.781, NPV 0.997, F1 score 0.745; testing dataset: sensitivity 0.756, specificity 0.995, PPV 0.532, NPV 0.998, F1 score 0.624; *Table 3*, *Figure 3E* and Figure S4).

Alternative hospital validation of the DL-based vertebra diagnostic system

We also applied the proposed vertebra diagnostic system to an alternative hospital validation dataset from another center (*Figure 4A,4B*). The alternative hospital validation dataset included 467 CT slices from 46 patients at Xuanwu Hospital (*Table 1*), and consisted of 645 manually annotated diseased vertebrae and 2,425 semi-automatically detected normal vertebrae (*Table 2*). The VCModule also exhibited good performance in diagnosing OVCF, OF, KD, and PS classes (sensitivity 0.891±0.111, specificity 0.989±0.021, PPV 0.902±0.103, NPV 0.997±0.002, F1 score 0.892±0.079, n=4; *Table 3* and *Figure 4B*). However, the VCModule exhibited relatively poor performance in SN diagnosis (sensitivity 0.213, specificity 1.000, PPV 0.842, NPV 0.994, F1 score 0.340, *Table 3* and *Figure 4B*). Our experimental results showed that the proposed vertebra diagnostic system can accurately and reliably diagnose OVCF, OF, KD, and surgery diseases.

Discussion

In this study, we established a deep-learning vertebra disease diagnosis system based on CT images that can accurately diagnose five vertebra diseases: OVCF, OF, SN, KD, and PS. We used a Faster R-CNN model to detect each vertebra in the CT image, and a subsequent multi-output DL model to achieve a vertebra-level diagnosis. The multi-output DL model is suitable even for cases with co-

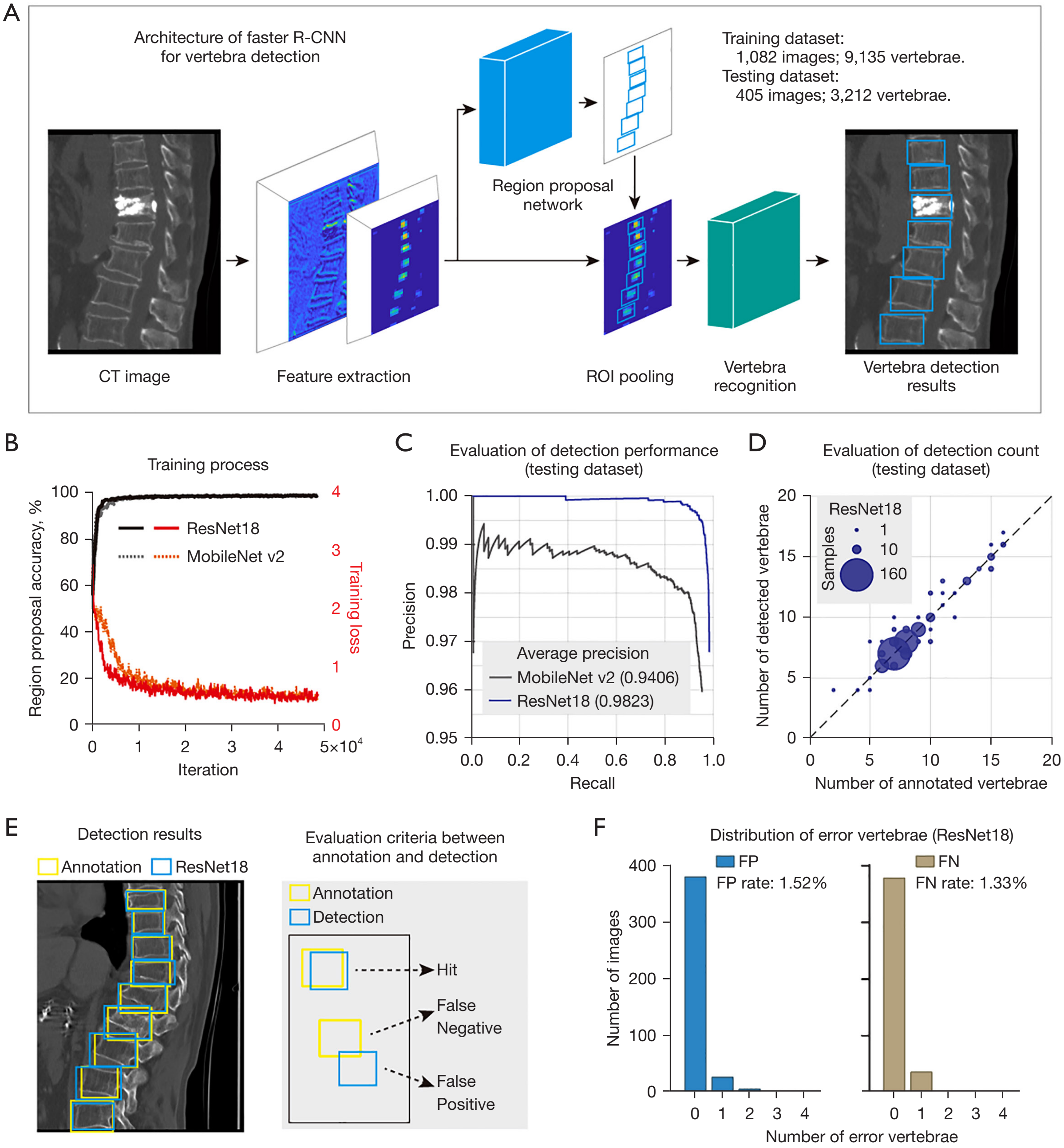

Figure 2 Automatic vertebra detection using the Faster R-CNN model. (A) Architecture of the Faster R-CNN for vertebrae. From left to right, clinical CT image, feature extraction operation implemented by CNN, region proposal network to generate Bbox candidates, ROI pooling to combine the features and Bbox candidates, vertebra recognition network, and vertebra detection results. (B) The plots of region proposal accuracy and loss versus iteration in the training process. Left axis, region proposal accuracy (black line). Right axis, training loss (red line). Solid line, ResNet18-based Faster R-CNN model. Dotted line, MobileNet v2-based Faster R-CNN model. (C) Evaluation of the detection performance in the testing dataset. The plot of the precision-recall curve. Gray, MobileNet v2-based model with a mAP of 0.9406. Blue, ResNet18-based model with a mAP of 0.9823. (D) Evaluation of the vertebra detection count in the testing dataset. Scatter plot of the number of detected vertebrae using the ResNet18-based model and the number of annotated vertebrae. The size of the scatter is proportional to the CT image sample size. (E) Spatial evaluation of vertebra detection. Left, visualization of the expert annotation and the ResNet18-based vertebra detection results. Right, an example of the evaluation criteria between annotation and detection. Three evaluation types: hit, false negative, and false positive. Yellow, annotated Bbox. Blue, detected Bbox. (F) Distribution of the error vertebrae in the ResNet18-based VDModule. Left, histogram plot of the number of images versus the number of FP vertebrae. The mean FP rate is 1.52%. Right, histogram plot of the number of images versus the number of FN vertebrae. The mean FN rate is 1.33%. R-CNN, region-based convolutional neural network; CT, computed tomography; CNN, convolutional neural network; ROI, region of interest; Bbox, bounding box; mAP, mean average precision; FP, false-positive; FN, false negative.

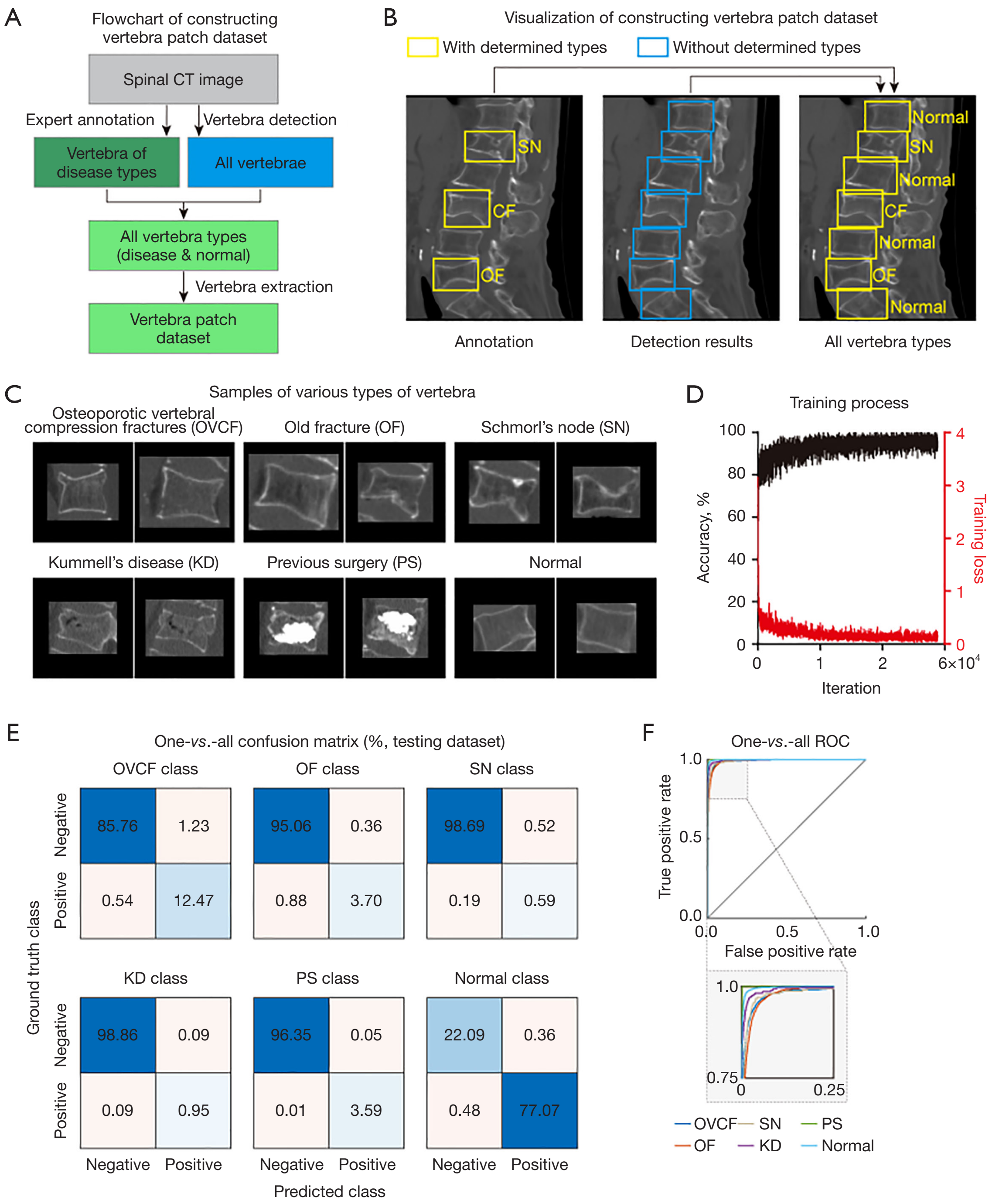

Figure 3 Intelligent vertebra classification using the multi-output DL model. (A) Flowchart showing the construction of the vertebra patch dataset to develop the vertebra classification models. Vertebrae with various disease types were annotated by experts. All vertebrae in the CT image were detected using the trained Faster R-CNN model. The normal vertebrae were acquired by subtracting the diseased vertebrae from all detected vertebrae. (B) Visualization of the method for constructing the vertebra patch dataset. Left, expert annotation of a vertebra with OVCF, a vertebra with OF, and a vertebra with SN. Middle, all vertebrae detected by the Faster R-CNN model. Right, diseased vertebrae and normal vertebrae. Yellow, vertebrae with determined types. Blue, vertebrae without determined types. (C) Sample of various types of vertebrae. The disease vertebrae include those with OVCF, OF, SN, KD, and PS. (D) The plots of training accuracy and loss versus iteration. Left axis, training accuracy (black line). Right axis, training loss (red line). (E) One-*vs.*-all confusion matrix plot for the ResNet50-based multi-output model in testing dataset. (F) One-*vs.*-all ROC curve. Insert plot, enlarged graph in the selected range. CT, computed tomography; SN, Schmorl's node; OVCF, osteoporotic vertebral compression fractures; OF, old fracture; KD, Kummell's disease; PS, previous surgery; ROC, receiver operating characteristic; DL, deep learning; R-CNN, region-based convolutional neural network.

Table 3 Quantitative evaluation of VCModule for six classes in the Luhe and Xuanwu Hospital datasets

Dataset	Class	Sensitivity	Specificity	PPV	NPV	F1 score
Lehe dataset (Train)	OVCF	0.984	0.991	0.939	0.998	0.961
	OF	0.886	0.998	0.954	0.994	0.919
	SN	0.713	0.998	0.781	0.997	0.745
	KD	0.917	1.000	0.963	0.999	0.939
	PS	0.993	1.000	1.000	1.000	0.997
Lehe dataset (Test)	OVCF	0.958	0.986	0.910	0.994	0.934
	OF	0.808	0.996	0.911	0.991	0.856
	SN	0.756	0.995	0.532	0.998	0.624
	KD	0.913	0.999	0.913	0.999	0.913
	PS	0.997	0.999	0.986	1.000	0.992
Xuanwu dataset (Test)	OVCF	0.980	0.957	0.775	0.997	0.866
	OF	0.741	0.997	0.860	0.994	0.796
	SN	0.213	1.000	0.842	0.994	0.340
	KD	0.874	1.000	0.982	0.998	0.925
	PS	0.970	1.000	0.991	0.999	0.980

VCModule, vertebra classification module; PPV, positive predictive value; NPV, negative predictive value; OVCF, osteoporotic vertebral compression fractures; OF, old fracture; SN, Schmorl's node; KD, Kummell's disease; PS, previous surgery.

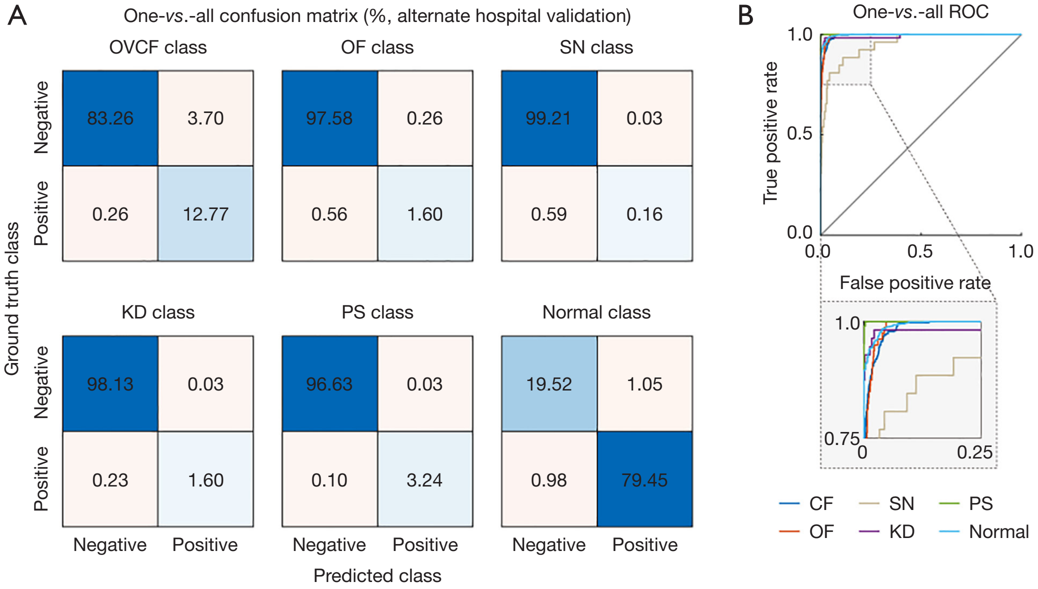

Figure 4 Alternative hospital validation of the DL-based vertebra diagnostic system. (A) One-*vs.*-all confusion matrix plot for the ResNet50-based multi-output model. (B) One-*vs.*-all ROC curve. Insert plot, enlarged graph in the selected range. OVCF, osteoporotic vertebral compression fractures; OF, old fracture; SN, Schmorl's node; KD, Kummell's disease; PS, previous surgery; ROC, receiver operating characteristic; DL, deep learning.

occurrence of multiple vertebra diseases.

Artificial intelligence-based diagnosis of vertebral diseases is gaining increasingly more attention. Prior literature reported several DL-based models in diagnosing OVCFs with the sensitivity of 0.54 to 0.85 and specificity of 0.92 to 0.96 (26,27). The diagnostic system proposed in this study was developed and validated in a larger dataset and achieved better performance. Furthermore, as most of the reported studies focused only on OVCFs (26,27), automatic diagnosis of multi-type vertebra diseases remains a challenging problem. Our proposed DL model achieved high performance in classifying four vertebra diseases: OVCF, OF, KD, and PS. The precise diagnosis of vertebral disease subtypes would play an important role in guiding subsequent surgery and treatment. We also achieved single vertebra-level diagnosis, which provides more accurate diagnostic information than a CT slice-level diagnosis (26,27). We used a Faster R-CNN model-based vertebra detection method to calculate the location of the vertebrae. In the literature, several studies have reported the on use of vertebral body segmentation to localize individual vertebrae (35,36). We also validated the generalization and reliability of the proposed system using an alternative hospital dataset acquired from another institute using a scanner and scanning parameters that differed from those used for the training and testing datasets. The proposed diagnostic system was based on two-dimensional (2D) images and 2D DL models. The adopted 2D models in this study worked fine in both vertebra detection and classification tasks. Besides, 2D models are usually easier to train and run faster than three-dimensional (3D) models. However, if 3D spatial features are key factors in a task, such as assessing vertebral morphology, the 3D DL models would be a better choice.

X-rays are also an important method for fracture diagnosis and have the advantages of fast speed, low cost, and low radiation intensity (37). Several X-ray-based DL models have been reported for fracture recognition (38,39). However, one study pointed out that lateral X-ray-based OVCF diagnostic models have diagnosis failure rates of as high as 20% for T5–T10 fracture (40). The occlusion by the diaphragm and lungs in lateral spine X-rays adversely affects the sensitivity and specificity of an X-ray-based DL model. Our CT-based DL model exhibited high performance in multiple vertebra disease diagnostic tasks, revealing that the combination of DL and CT would be a promising method for vertebra diagnosis.

The proposed DL-based diagnostic system for multi-type vertebra diseases used supervised learning. Although this method requires lots of prior diagnostic results as training data, it has the advantages of high accuracy and reliability and is suitable for clinical diagnosis. After the diagnostic system is established, the system can be integrated into clinical workflow in several ways. First, the proposed DL system can diagnose lots of samples very quickly, and it can even realize real-time diagnosis by interfacing with imaging instruments. The proposed DL system is especially suitable for emergencies or occasions of lacking medical resources. Second, the system can be retrained in real clinical workflow to further improve its performance. One possible way is that experts reconfirm difficult cases based on the system's diagnostic results and correct the cases that are incorrectly diagnosed by the DL system. These corrected cases are then incorporated into the training data and retrain the diagnostic models to continuously improve their performance.

This study had several limitations. First, the training and testing datasets had class imbalance problem. The number of SN and KD samples was much smaller than other vertebra types, such as OVCF, OF, PS, and the normal. Although random over-sampling was adopted to overcome this problem, the resampling operation would introduce bias in the model. Second, the proposed system performed diagnosis at the vertebral level and slice level rather than at the patient level. The proposed system can be applied to determine whether a patient has certain vertebrae diseases (such as OVCF). For example, if one and more slice of a target vertebra is diagnosed with OVCF, the system can alert the radiologist that the patient is at higher risk of OVCF. Radiologists can focus on these high-risk vertebrae quickly and spend less time examining low-risk vertebrae, thus improving work efficiency. However, the current system does not recognize the vertebra position and it does not assess the overall condition of all vertebrae, leading to the inability in performing more complex tasks such as assessing the spine's health condition and predicting re-fracture probability. Third, the SN diagnosis performance was relatively poor. In both the training and testing datasets, the proposed system exhibited high specificity and NPV but poor sensitivity and PPV. These results suggested that the model was unfit for the SN diagnostic task and did not capture the key features of SN. There may be two main reasons. First, the number of SN samples was too few. Second, the key features of SN were closer to OVCF and OF, leading to high difficulty in recognizing SN. Notably, the number of KD was even less than SN, but DL models recognized KD accurately on both training and testing datasets. The main reason for this phenomenon is that KD has obvious features that significantly differentiate it from other types of vertebral diseases. Establishing a larger-scale and high-quality SN dataset as well as adopting more powerful DL architectures would be possible ways to achieve better performance in SN diagnosis. In future research, we plan to increase the size of the datasets and build multicenter datasets that can more accurately validate the reliability and generalization ability of the system. Furthermore, we would directly compare the model's performance with that of human experts to outline the importance of the DL models, revealing the potential value of the diagnostic system in real-world clinical applications.

Conclusions

In conclusion, the proposed DL system can accurately diagnose four vertebra diseases: OVCF, OF, KD, and PS. To the best of our knowledge, no other DL system has been reported to perform this refined multi-type diagnosis of vertebra diseases. The proposed CT-based DL system provides an innovative method for multi-type vertebral disease diagnosis and has strong potential to facilitate accurate and rapid diagnosis of vertebral diseases.

Acknowledgments

Funding: This study was supported by Tongzhou District Health Development Research Project (No. KJ2023CX032).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-685/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-685/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This retrospective study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was performed in accordance with relevant guidelines and regulations, and approved by the Institutional Review Board of Beijing Luhe Hospital, Capital Medical University (No. 2023-LHKY-022-02). Written informed consent was taken from all individual participants.

References

1. Silverman SL. The clinical consequences of vertebral compression fracture. Bone 1992;13 Suppl 2:S27-31.
2. Kanterewicz E, Puigoriol E, Rodríguez Cros JR, Peris P. Prevalent vertebral fractures and minor vertebral deformities analyzed by vertebral fracture assessment (VFA) increases the risk of incident fractures in postmenopausal women: the FRODOS study. Osteoporos Int 2019;30:2141-9.
3. Lane NE. Epidemiology, etiology, and diagnosis of osteoporosis. Am J Obstet Gynecol 2006;194:S3-11.
4. Ballane G, Cauley JA, Luckey MM, El-Hajj Fuleihan G. Worldwide prevalence and incidence of osteoporotic vertebral fractures. Osteoporos Int 2017;28:1531-42.
5. Wang O, Hu Y, Gong S, Xue Q, Deng Z, Wang L, Liu H, Tang H, Guo X, Chen J, Jia X, Xu Y, Lan L, Lei C, Dong H, Yuan G, Fu Q, Wei Y, Xia W, Xu L. A survey of outcomes and management of patients post fragility fractures in China. Osteoporos Int 2015;26:2631-40.
6. Edidin AA, Ong KL, Lau E, Kurtz SM. Morbidity and Mortality After Vertebral Fractures: Comparison of Vertebral Augmentation and Nonoperative Management in the Medicare Population. Spine (Phila Pa 1976) 2015;40:1228-41.
7. Marques A, Ferreira RJ, Santos E, Loza E, Carmona L, da Silva JA. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. Ann Rheum Dis 2015;74:1958-67.
8. Cummings SR, Melton LJ. Epidemiology and outcomes of osteoporotic fractures. Lancet 2002;359:1761-7.
9. Wong CC, McGirt MJ. Vertebral compression fractures: a review of current management and multimodal therapy. J

Multidiscip Healthc 2013;6:205-14.

10. Delmas PD, van de Langerijt L, Watts NB, Eastell R, Genant H, Grauer A, Cahall DL; . Underdiagnosis of vertebral fractures is a worldwide problem: the IMPACT study. J Bone Miner Res 2005;20:557-63.
11. Malik AT, Retchin S, Phillips FM, Xu W, Peters K, Yu E, Khan SN. Declining trend in osteoporosis management and screening following vertebral compression fractures - a national analysis of commercial insurance and medicare advantage beneficiaries. Spine J 2020;20:538-46.
12. Puisto V, Heliövaara M, Impivaara O, Jalanko T, Kröger H, Knekt P, Aromaa A, Rissanen H, Helenius I. Severity of vertebral fracture and risk of hip fracture: a nested case-control study. Osteoporos Int 2011;22:63-8.
13. Expert Panels on Neurological Imaging, Interventional Radiology, and Musculoskeletal Imaging; Shah LM, Jennings JW, Kirsch CFE, Hohenwalter EJ, Beaman FD, Cassidy RC, Johnson MM, Kendi AT, Lo SS, Reitman C, Sahgal A, Scheidt MJ, Schramm K, Wessell DE, Kransdorf MJ, Lorenz JM, Bykowski J. ACR Appropriateness Criteria® Management of Vertebral Compression Fractures. J Am Coll Radiol 2018;15:S347-S364.
14. Jung HS, Jee WH, McCauley TR, Ha KY, Choi KH. Discrimination of metastatic from acute osteoporotic compression spinal fractures with MR imaging. Radiographics 2003;23:179-87.
15. Piazzolla A, Solarino G, Lamartina C, De Giorgi S, Bizzoca D, Berjano P, Garofalo N, Setti S, Dicuonzo F, Moretti B. Vertebral Bone Marrow Edema (VBME) in Conservatively Treated Acute Vertebral Compression Fractures (VCFs): Evolution and Clinical Correlations. Spine (Phila Pa 1976) 2015;40:E842-8.
16. Liao H, Mesfin A, Luo J. Joint Vertebrae Identification and Localization in Spinal CT Images by Combining Short- and Long-Range Contextual Information. IEEE Trans Med Imaging 2018;37:1266-75.
17. VandenBerg J, Cullison K, Fowler SA, Parsons MS, McAndrew CM, Carpenter CR. Blunt Thoracolumbar-Spine Trauma Evaluation in the Emergency Department: A Meta-Analysis of Diagnostic Accuracy for History, Physical Examination, and Imaging. J Emerg Med 2019;56:153-65.
18. Kaup M, Wichmann JL, Scholtz JE, Beeres M, Kromen W, Albrecht MH, Lehnert T, Boettcher M, Vogl TJ, Bauer RW. Dual-Energy CT-based Display of Bone Marrow Edema in Osteoporotic Vertebral Compression Fractures: Impact on Diagnostic Accuracy of Radiologists with Varying Levels of Experience in Correlation to MR Imaging. Radiology 2016;280:510-9.
19. Mao H, Zou J, Geng D, Zhu X, Zhu M, Jiang W, Yang H. Osteoporotic vertebral fractures without compression: key factors of diagnosis and initial outcome of treatment with cement augmentation. Neuroradiology 2012;54:1137-43.
20. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, Sköldenberg O, Gordon M. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88:581-6.
21. Su F, Li J, Zhao X, Wang B, Hu Y, Sun Y, Ji J. Interpretable tumor differentiation grade and microsatellite instability recognition in gastric cancer using deep learning. Lab Invest 2022;102:641-9.
22. Hu Y, Su F, Dong K, Wang X, Zhao X, Jiang Y, Li J, Ji J, Sun Y. Deep learning system for lymph node quantification and metastatic cancer identification from whole-slide pathology images. Gastric Cancer 2021;24:868-77.
23. Qu B, Cao J, Qian C, Wu J, Lin J, Wang L, Ou-Yang L, Chen Y, Yan L, Hong Q, Zheng G, Qu X. Current development and prospects of deep learning in spine image analysis: a literature review. Quant Imaging Med Surg 2022;12:3454-79.
24. Su F, Cheng Y, Chang L, Wang L, Huang G, Yuan P, Zhang C, Ma Y. Annotation-free glioma grading from pathological images using ensemble deep learning. Heliyon 2023;9:e14654.
25. Zhao W, Shen L, Islam MT, Qin W, Zhang Z, Liang X, Zhang G, Xu S, Li X. Artificial intelligence in image-guided radiotherapy: a review of treatment target localization. Quant Imaging Med Surg 2021;11:4881-94.
26. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med 2018;98:8-15.
27. Kolanu N, Silverstone EJ, Ho BH, Pham H, Hansen A, Pauley E, Quirk AR, Sweeney SC, Center JR, Pocock NA. Clinical Utility of Computer-Aided Diagnosis of Vertebral Fractures From Computed Tomography Images. J Bone Miner Res 2020;35:2307-12.
28. Kim YC, Kim YH, Ha KY. Pathomechanism of intravertebral clefts in osteoporotic compression fractures of the spine. Spine J 2014;14:659-66.
29. Zhu J, Yang S, Yang Y, Yao T, Liu G, Fan S, Zhao H, Cui F, Wang X, Jiang G, Fang X. Modified poly(methyl methacrylate) bone cement in the treatment of Kümmell disease. Regen Biomater 2021;8:rbaa051.
30. Chen L, Dong R, Gu Y, Feng Y. Comparison between Balloon Kyphoplasty and Short Segmental Fixation Combined with Vertebroplasty in the Treatment of

Kümmell's Disease. Pain Physician 2015;18:373-81.

31. Resnick D, Niwayama G. Intravertebral disk herniations: cartilaginous (Schmorl's) nodes. Radiology 1978;126:57-65.
32. Abbas J, Hamoud K, Peled N, Hershkovitz I. Lumbar Schmorl's Nodes and Their Correlation with Spine Configuration and Degeneration. Biomed Res Int 2018;2018:1574020.
33. Griffith JF. Identifying osteoporotic vertebral fracture. Quant Imaging Med Surg 2015;5:592-602.
34. Su F, Sun Y, Hu Y, Yuan P, Wang X, Wang Q, Li J, Ji JF. Development and validation of a deep learning system for ascites cytopathology interpretation. Gastric Cancer 2020;23:1041-50.
35. Schmidt D, Ulén J, Enqvist O, Persson E, Trägårdh E, Leander P, Edenbrandt L. Deep learning takes the pain out of back breaking work - Automatic vertebral segmentation and attenuation measurement for osteoporosis. Clin Imaging 2022;81:54-9.
36. Suri A, Jones BC, Ng G, Anabaraonye N, Beyrer P, Domi A, Choi G, Tang S, Terry A, Leichner T, Fathali I, Bastin N, Chesnais H, Rajapakse CS. A deep learning system for automated, multi-modality 2D segmentation of vertebral bodies and intervertebral discs. Bone 2021;149:115972.
37. Tozakidou M, Reisinger C, Harder D, Lieb J, Szucs-Farkas Z, Müller-Gerbl M, Studler U, Schindera S, Hirschmann A. Systematic Radiation Dose Reduction in Cervical Spine CT of Human Cadaveric Specimens: How Low Can We Go? AJNR Am J Neuroradiol 2018;39:385-91.
38. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, Chung IF, Liao CH. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29:5469-77.
39. 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;479:1598-612.
40. Guglielmi G, Palmieri F, Placentino MG, D'Errico F, Stoppino LP. Assessment of osteoporotic vertebral fractures using specialized workflow software for 6-point morphometry. Eur J Radiol 2009;70:142-8.

Cite this article as: Wang Y, Su F, Lu Q, Zhang W, Liu T, Tao Y, Fu S, Cui L, Lu SB, Chen X, Shi Z. Deep learning-driven diagnosis of multi-type vertebra diseases based on computed tomography images. Quant Imaging Med Surg 2024;14(1):800-813. doi: 10.21037/qims-23-685

Supplementary

Appendix 1

CT scanner and imaging parameters

At Luhe Hospital, the CT images were acquired in an axial orientation using a Philips Ingenuity Core 128 CT scanner at 120 kV, 300 mAs, and 1 mm thickness. The acquired images were reconstructed using a Philips Portal Workstation in sagittal view (3 mm thickness) in the digital imaging and communications in medicine (DICOM) format and downloaded to Hina MIIS-RIS PACS system. At Xuanwu Hospital, the CT images were acquired in axial orientation using a GE Revolution CT scanner at 120 kV, 200-500 mAs, and 0.625 mm thickness. The acquired images were reconstructed using a GE AW4.7 Workstation to sagittal view (2 mm thickness) in DICOM format and downloaded to a UniWeb Viewer PACS system.

We further used a Hina MIIS-RIS at Luhe Hospital and a UniWeb Viewer system at the Xuanwu Hospital, and we converted the CT slices from the DICOM format to the JPEG format. The format conversion was achieved by linear mapping of the values in the DICOM image to a minimum of 0 and a maximum of 255. Converting the DICOM image to a JPEG image led to a perceptible loss in image quality, but this quality loss had little effect on the ability of the surgeons to make a correct diagnosis.

Annotation of dataset

Surgeons implemented annotations on JPEG image of each CT slice using the LabelMe tool (version 5.0.1). The injured vertebrae annotation process consisted of three steps: independent annotation, consistency checking, and consulted co-annotation (*Figure S1*,*S2*). First, three spine surgeons implemented diagnoses and annotations for every CT slice independently. Second, we implemented consistency checking for all annotations. We compared the annotations on each CT slice from the three surgeons to find inconsistent annotations. Finally, for the inconsistent annotations, all three surgeons rechecked the patient's spinal CT images and corresponding MRI T1 and T2 images together to reach consensus.

Parameter settings for vertebra detection module

Vertebra detection module (VDModule) was developed to detect all vertebrae in teach CT slice. For the object detection task, the Bboxes are rectangular and are sensitive to random rotation operation. Rotation augmentation in an arbitrary degree would affect the precision of Bboxes and make the ground truth less reliable, thereby likely hurting model performance. We ensured the reliability of the Bboxes after augmentation by adopting several specific operations, including horizontal and vertical flipping, rotating 90 degrees, or a multiple of 90 degrees. We enlarged the image and Bbox datasets eight-fold using offline data augmentation (*Figure S3*).

Other parameter settings were set as follows: batch size 3; maximum epoch 5; learning rate 10^{-3}; and stochastic gradient descent with momentum (SGDM) optimizer. The DL architectures and experiments were implemented on a computer with MATLAB 2021a and configured with an Nvidia GeForce GTX 1080 Ti GPU with 11 GB of memory.

Calculation of Bboxes for normal category vertebrae

In the annotation process of the vertebra classification task, only the vertebrae with injury were annotated in each image. The remaining vertebrae were the "normal" category. We automatically calculated the Bboxes of the normal category vertebrae. First, the Bboxes of all vertebrae were determined using the developed VDModule for each CT slice. Second, we abandoned the detected Bboxes that largely overlapped with at least one annotated disease Bbox (IoM ≥0.5). The remaining annotated Bboxes that had little or no spatial overlap with the disease Bboxes were deemed normal category vertebrae.

Development of the vertebra extraction module

The vertebra extraction module (VEModule) was used to construct the vertebra image patch dataset for the vertebra classification task. Each Bbox of a CT slice underwent three steps in the image patch extraction process. First, the Bbox was augmented five-fold through translation and scaling. The horizontal and vertical translation range was [-10 to 10] pixels. The value of the scaling range was [0.9–1.1]. Second, the CT image was cropped using each Bbox and the image size was modified to 196×196 pixels using zero-padding and center-crop operations. Third, the image patch was rescaled to 224×224 pixels to fit the subsequent DL models. For the training dataset, all augmented image patches were used for training. For the testing dataset, the average diagnostic results of the augmented image patches were used as the final diagnostic results.

Parameter settings for vertebra classification module

We developed vertebra classification module (VCModule) to classify six vertebra categories. We used random over-sampling and under-sampling strategies to solve the class imbalance problem, and used data augmentation techniques to solve the overfitting problem.

In the training dataset, the samples for different vertebra diseases were highly imbalanced. We used random over-sampling and random under-sampling strategies to establish a more balanced training dataset for various vertebra categories. Over-sampling can lead to an overfitting problem, especially for classes with few original samples. Hence, we used data augmentation techniques to solve the overfitting problem. The data augmentation consisted of multiple image processing operations, including image rotation, image translation, noise addition, and brightness and contrast modification. Other parameter settings were as follows: maximum epoch, 15; batch size, 64; learning rate, 10-2; learning rate decays with a ratio of 0.2 every 5 epochs; and SGDM optimizer.

Table S1 Distribution of diseased thoracic and lumbar vertebrae

Vertebra location	Luhe Hospital cohort: patient count (percentage)	Xuanwu Hospital cohort: patient count (percentage)
T1	0 (0.0%)	0 (0.0%)
T2	0 (0.0%)	0 (0.0%)
T3	0 (0.0%)	0 (0.0%)
T4	1 (0.1%)	0 (0.0%)
T5	9 (0.9%)	0 (0.0%)
T6	34 (3.2%)	0 (0.0%)
T7	41 (3.9%)	1 (2.2%)
T8	49 (4.7%)	3 (6.5%)
T9	39 (3.7%)	2 (4.3%)
T10	39 (3.7%)	1 (2.2%)
T11	86 (8.2%)	3 (6.5%)
T12	282 (26.8%)	11 (23.9%)
L1	285 (27.1%)	14 (30.4%)
L2	148 (14.1%)	7 (15.2%)
L3	107 (10.2%)	11 (23.9%)
L4	84 (8.0%)	3 (6.5%)
L5	24 (2.3%)	1 (2.2%)

T, thoracic vertebra; L, lumbar vertebra.

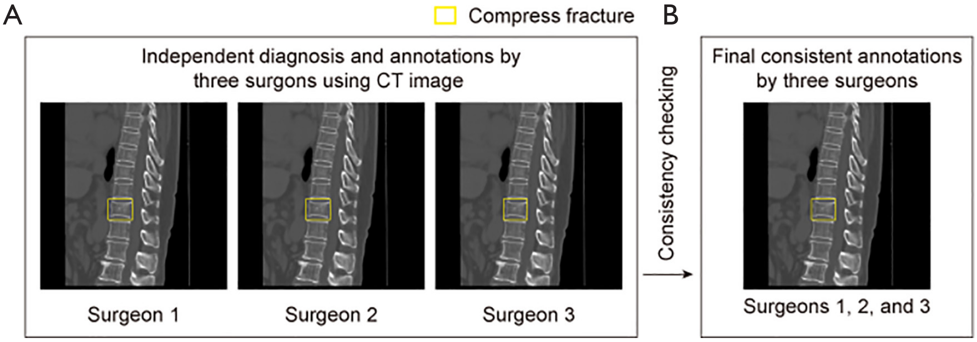
A
Compress fracture
B
Independent diagnosis and annotations by three surgons using CT image
Surgeon 1
Surgeon 2
Surgeon 3
Consistency checking
Final consistent annotations by three surgeons
Surgeons 1, 2, and 3

Figure S1 Annotation process for vertebrae with consistent initial diagnoses. (A) Consistent diagnosis and annotations on a computer tomography (CT) slice by three surgeons independently. (B) Final consistent annotations by the three surgeons.

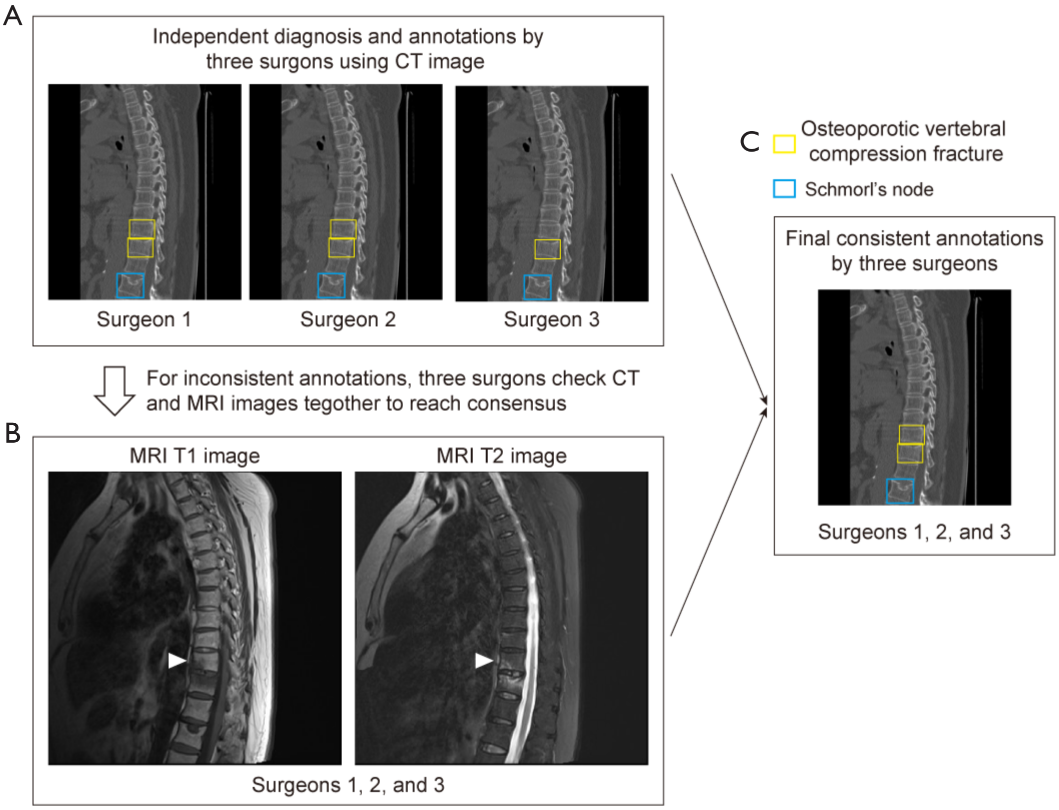
A
Independent diagnosis and annotations by three surgons using CT image
Surgeon 1
Surgeon 2
Surgeon 3
For inconsistent annotations, three surgons check CT and MRI images tegother to reach consensus
B
MRI T1 image
MRI T2 image
Surgeons 1, 2, and 3
C
Osteoporotic vertebral compression fracture
Schmorl's node
Final consistent annotations by three surgeons
Surgeons 1, 2, and 3

Figure S2 Annotation process for vertebrae with inconsistent initial diagnoses. (A) Inconsistent diagnosis and annotations on a computer tomography (CT) slice by three surgeons independently. All surgeons diagnosed thoracic vertebra T11 and lumbar vertebra L1 as osteoporotic vertebral compression fracture (OVCF) and Schmorl's node (SN), respectively. Surgeons 1 and 2 diagnosed thoracic vertebra T10 as OVCF, but surgeon 3 diagnosed T10 as Normal. (B) Consulted co-annotation by the three surgeons for inconsistent annotations based on CT images and the corresponding MRI T1 and T2 images. The arrows indicate that T10 has obvious OVCF characteristics on MRI T1 and T2 images. (C) Final consistent annotations by the three surgeons. T10, T11 and L1 were diagnosed as OVCF, OVCF, and SN, respectively.

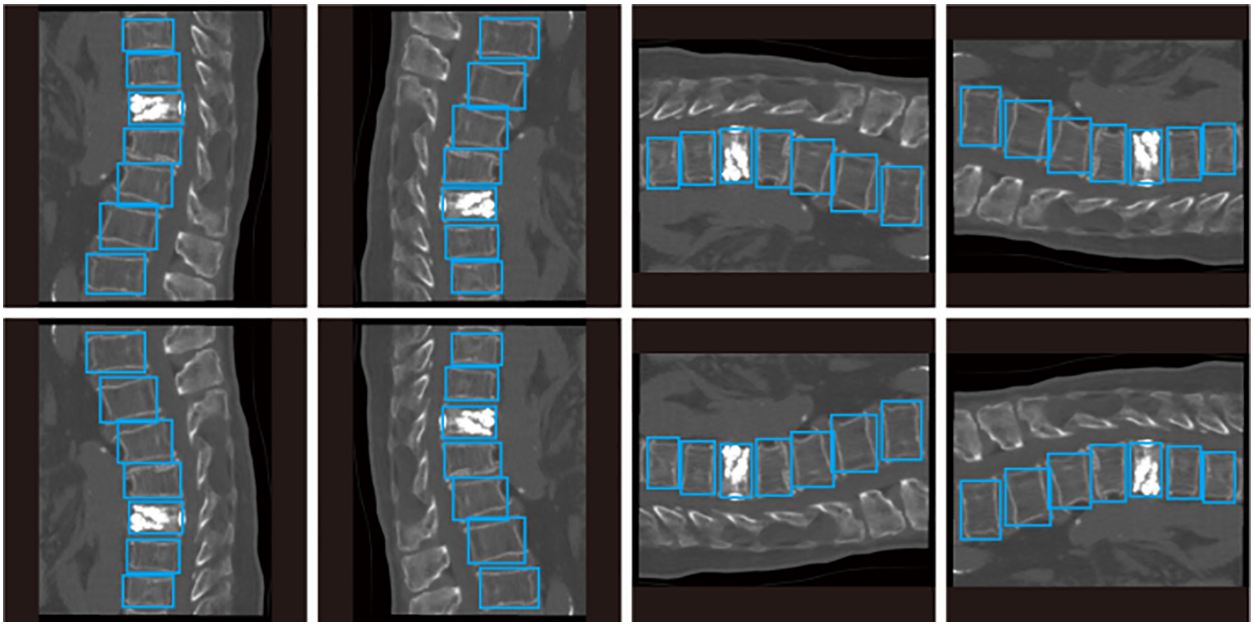

Figure S3 Eight types of augmentation operations used for both image and bounding boxes. Each blue rectangle represents one annotated or detected vertebra.

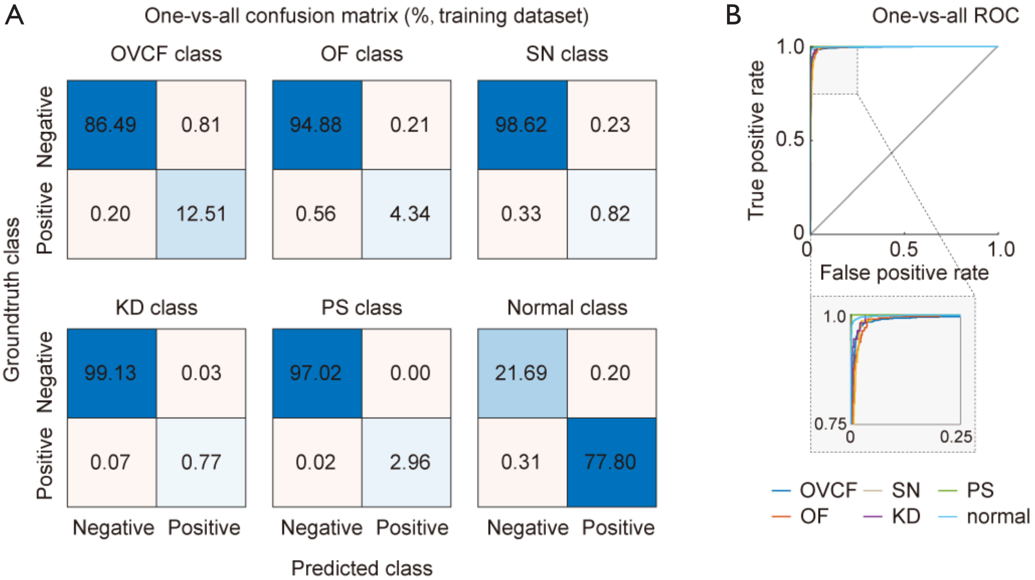

Figure S4 Performance of the DL-based vertebra diagnostic system in the training dataset. (A) One-*vs.*-all confusion matrix plot for ResNet50-based multi-output model. (B) One-*vs.*-all ROC curve. Insert plot, enlarged graph in the selected range.