Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A machine learning based quantification system for automated diagnosis of lumbar spondylolisthesis on spinal X-rays

Shanshan Liu[a,b,c,1], Chenyi Guo[d,1], Yuting Zhao[d,1], Cheng Zhang[a,b,c,1], Lihao Yue[e], Ruijie Yao[d], Qifeng Lan[e], Xingyu Zhou[e], Bo Zhao[e], Ji Wu[d,***,2], Weishi Li[a,b,c,2,**], Nanfang Xu[a,b,c,*,2]

[a] *Department of Orthopaedics, Peking University Third Hospital, Beijing, China*
[b] *Engineering Research Center of Bone and Joint Precision Medicine, Ministry of Education, Beijing, China*
[c] *Beijing Key Laboratory of Spinal Disease Research, Beijing, China*
[d] *Department of Electronic Engineering, Tsinghua University, Beijing, China*
[e] *Peking University Health Science Center, Beijing, China*

ARTICLE INFO

Keywords:
Spondylolisthesis
Machine learning
Automated diagnosis
Prime sample attention

ABSTRACT

The automated diagnosis of lumbar spondylolisthesis lacks standardized criteria and the diagnostic of lumbar spondylolisthesis itself inherently relies on the subjective judgment of experts, resulting in a lack of standardized criteria. The objective of this study is to develop a novel, fully automated diagnostic system for lumbar spondylolisthesis. A two-stage system was developed, consisting of a Mask R-CNN with Prime Sample Attention (PISA) for vertebral segmentation in the first stage and a Light Gradient Boosting Machine (LGBM) for spondylolisthesis diagnosis in the second stage. The training data was developed by a total of 936 X-ray images including neutral, extension, and flexion lateral radiographs retrospectively gathered from 312 patients diagnosed with lumbar spondylolisthesis between January 2021 and March 2022. From a model perspective, there were a total of 4680 vertebrae, of which 1056 (22.6 %) were spondylolisthesis and the rest were normal. The Bbox mAP50, Bbox mAP75, Segm mAP50, and Segm mAP75 of Mask R-CNN with PISA were 0.9852, 0.9179, 0.9741, and 0.8957, respectively. The Accuracy, AUC, Recall, Precision, and F1-score of LGBM were 0.9660, 0.9843, 0.9020, 0.9020, and 0.9020, respectively. This study presents a robust system for the diagnosis of lumbar spondylolisthesis, providing accurate detection, classification, and localization of lumbar spondylolisthesis.

1. Introduction

Lumbar spondylolisthesis stands as one of the most prevailing lumbar disorders [1–4]. The stability of the lumbar spine hinges upon

* Corresponding author. Peking University Third Hospital, 49 North Garden Road, Beijing, 100191, China.
** Corresponding author. Peking University Third Hospital, 49 North Garden Road, Beijing 100191, China.
*** Corresponding author. Tsinghua University, 30 Shuangqing Road, Beijing, 100084, China.
E-mail addresses: wuji_ee@mail.tsinghua.edu.cn (J. Wu), puh3liweishi@163.com (W. Li), xunanfang@foxmail.com (N. Xu).

[1] Shanshan Liu, Chenyi Guo, Yuting Zhao, and Cheng Zhang are **co-first authors**.
[2] Nanfang Xu, Weishi Li, and Ji Wu are **co-corresponding authors**.

https://doi.org/10.1016/j.heliyon.2024.e37418
Received 27 August 2024; Received in revised form 3 September 2024; Accepted 3 September 2024
Available online 4 September 2024

the integrity of the vertebral arch, intervertebral disc, ligaments, and surrounding soft tissues. Disruptions of these components due to degenerative processes can culminate in vertebral slippage, with relative displacement observed between the adjacent vertebrae. The most suitable non-invasive examination method for detecting spondylolisthesis is the standing lumbar lateral radiograph based on relevant guidelines [5,6]. Although methods for assessing patients with a confirmed diagnosis of spondylolisthesis have been widely applied in clinical practice, such as the Meyerding classification for evaluating the degree of slippage and the Wiltse classification for assessing its etiology [7,8], determining the presence of spondylolisthesis in a broader population that includes a large number of healthy individuals and asymptomatic cases remains an unresolved issue. Some subtle slippages are challenging to differentiate between normal variations and those indicative of spondylolisthesis, complicating the diagnostic process. Some studies have indicated that inter-observer and intra-observer variability in lumbar spine diagnosis can reach up to 15 % [9]. The decision for whether or not relative vertebral displacement exists is largely subjective (even by experts) and lacks any standardized and quantifiable distance or angular criteria [1].

Over recent years, deep learning models have demonstrated remarkable proficiency in the detection and segmentation of specific targets in spinal X-ray images [10–22]. Automatic detection and segmentation of the entire spine or specific regions such as the cervical and lumbar spine have already been achieved [14,16,18,19]. Furthermore, leveraging key vertebral landmarks obtained from deep learning models, some researchers are focusing on the automatic measurement of spine-related angles and distances [11–13,15]. This approach significantly reduces time and cost, and these automatically measured angles and distances can serve as key features for diagnosing or assessing spinal diseases such as scoliosis [15] and sagittal spinal imbalance [22]. Trinh et al. [21], manually defined geometric features between lumbar vertebrae for the detection of lumbar spondylolisthesis, building upon key vertebral landmarks identified by computer vision models. Nevertheless, manually defined features struggle to capture the rich information between vertebral bodies, leaving considerable room for improvement in diagnostic accuracy.

The objective of our study is to develop an automated diagnostic system using two-stage machine learning models for lumbar spondylolisthesis. In the first stage, the system employs an instance segmentation model, commonly used in the medical imaging domain, to annotate lateral lumbar X-rays. It then extracts the coordinates of key vertebral landmarks and performs coordinate refinement and feature extraction through a post-processing module. In the second stage, the diagnosis of lumbar spondylolisthesis is carried out based on the Light Gradient Boosting Machine (LGBM) model, utilizing the key point coordinates and their geometric features in lateral X-ray images obtained in neutral, extension, and flexion positions. The system also returns the specific lumbar vertebrae exhibiting spondylolisthesis on each X-ray image. Overall, this system is capable of automatically providing a diagnosis of lumbar spondylolisthesis and localizing the specific vertebrae affected by spondylolisthesis from the initial X-ray evaluation.

2. Methods

The diagnostic method for lumbar spondylolisthesis that we propose is illustrated in Fig. 1. Initially, we acquire X-ray images from three different perspectives: neutral, lordotic, and kyphotic. The image processing procedure comprises two main stages.

The image processing is divided into two independent parts. On one hand, the locations of the lumbar vertebrae in the images are extracted using an instance segmentation model, which subsequently provides the vertebral coordinates and geometric features. On the other hand, due to the different facial orientations during X-ray image acquisition, we determine the facial orientation by comparing the areas of the left and right sides after threshold segmentation. Using this orientation information, we unify the coordinates of all X-ray images to the right side.

After obtaining the coordinates of each lumbar vertebral body, we calculate the “Slip ratio” based on the geometric relationship between the lumbar vertebrae. We then combine the length, width, slope, and other information of the lumbar vertebrae from different view images, along with the lumbar vertebral numbering information, to construct a 28-dimensional feature vector for each lumbar

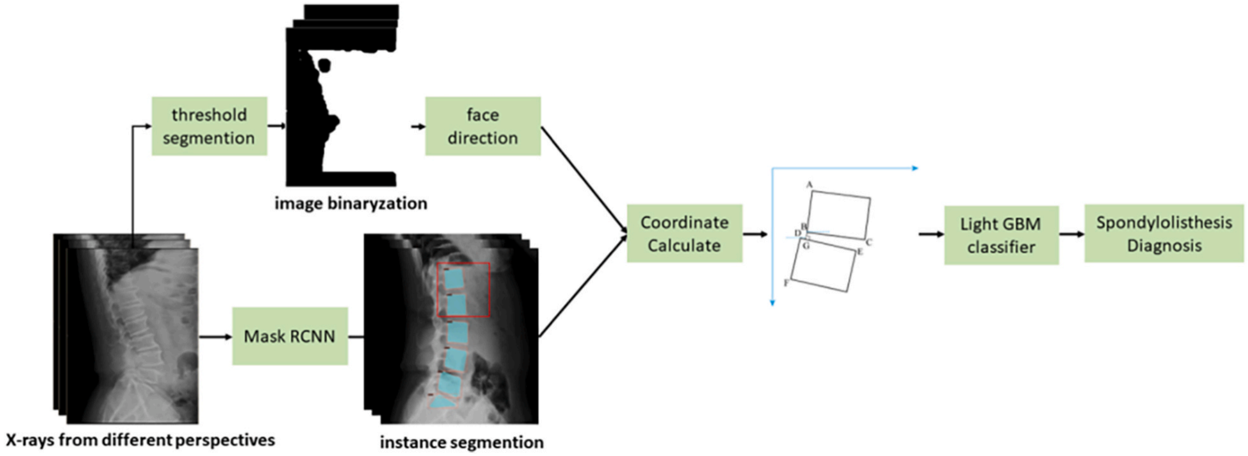

Fig. 1. The proposed architecture for lumbar spondylolisthesis diagnosis.

vertebra. Using these feature vectors, we train a Light GBM model for diagnosing lumbar spondylolisthesis, treating each lumbar vertebra as a unit.

2.1. Study design and participants

This research, aimed at developing and validating a diagnostic model, was designed as a single-center retrospective medical study. Approval for the study was obtained from the Peking University Third Hospital Institutional Review Board in March 3rd, 2022 (S2022290). Informed consent was waived due to the retrospective nature of the investigation.

X-ray images from patients diagnosed with lumbar spondylolisthesis between January 2021 and March 2022 were retrospectively gathered. Inclusion criteria were as follows: (1) confirmed diagnosis of spondylolisthesis diagnosis according to surgical notes, (2) availability of preoperative neutral, lordotic, and kyphotic lateral lumbar X-ray images, and (3) each X-ray image containing a minimum of 6 vertebral segments from L1 to S1. Exclusion criteria included patients who had a previous history of lumbar surgery, any other forms of spinal deformity, or incomplete pre-operative X-ray studies.

2.2. Dataset

In the first phase of the study, the sample comprised X-ray images, totaling 936 X-ray images (paired in three positions) from 312 patients, including 88 males (28.2 %) and 224 females (71.8 %). During the division of the training, validation, and test sets, it was ensured that the three images of the same patient were included in the same set. In the second phase, the sample consisted of a 28-dimensional feature vector constructed for each lumbar vertebra in the paired three positions. Since each set of paired X-ray images contained five samples (L1 to L5 vertebrae), there were a total of 1560 samples, of which 352 (22.6 %) were positive samples indicating spondylolisthesis, and the remaining were negative samples representing normal vertebrae. Additionally, both forms of spondylolisthesis (anterior slipping and the less common posterior slipping) were defined as positive samples. The datasets for both phases were divided into training (70 %), validation (10 %), and test sets (20 %). The baseline characteristics are displayed in Table 1.

2.3. Development of instance segmentation model

To achieve the diagnosis of lumbar spondylolisthesis, it is essential to determine the coordinates of each lumbar vertebra to obtain the geometric relationships between the vertebral bodies, which are crucial for identifying spondylolisthesis. The instance segmentation model serves as the first-stage model in the automated diagnostic system for lumbar spondylolisthesis. Its task is to extract the coordinates of each lumbar vertebra from the original X-ray images, thereby facilitating the subsequent second-stage spondylolisthesis diagnosis based on these coordinates.

To accurately obtain the coordinates of each vertebral body, performing instance segmentation on the vertebrae is a relatively intuitive approach. Instance segmentation is a complex task that requires the simultaneous completion of both object detection and semantic segmentation. There are two branches of instance segmentation algorithm, one is top-down segmentation based on object detection and detection, and the other is bottom-up instance segmentation based on semantic segmentation. Due to the fact that the content we annotate on the dataset is the coordinates of the box for object detection, it is suitable to use a top-down instance segmentation based on object detection. Mask R-CNN [23], which is based on the prototype of Faster R-CNN [24], adds branches for segmentation tasks to achieve instance segmentation, which naturally aligns with our data annotation format and tasks. Similar methods have proven effective in diagnosing other spinal diseases [25]. Therefore, we first adopt Mask R-CNN to perform instance segmentation. In Mask R-CNN, Each Region of Interest (ROI) predicts a segmentation mask using a small Fully Convolutional Network (FCN) [26] to retain spatial structure information. This allows instance segmentation to be achieved based on object detection.

Mask R-CNN is a seminal instance segmentation model composed of several key components [23]. The backbone feature extraction network, typically a pre-trained convolutional neural network like ResNet or VGG, is responsible for encoding visual features from the input image. The Region Proposal Network (RPN) generates candidate object regions of interest (RoIs) [24]. The RoI Align module then aligns these variable-sized RoIs to a fixed-size feature representation [27].

In the Mask R-CNN model, the multi task loss for each sampled RoI is defined as L, which consists of classification loss L_{cls}, regression loss L_{box}, and mask loss L_{mask}, respectively. Among them, L_{cls} and L_{box} are exactly the same as the definition of Faster R-CNN. The mask branch of Mask R-CNN has a K × M dimensional output for each RoI, encoding K binary masks with a resolution of m × m, where K represents the total number of categories. For this, apply sigmoid to each pixel and L_{mask} is defined as the average binary cross loss.

Table 1
Baseline characteristics.

Lumbar segment	Numbers of slipping vertebrae	Ratio of anterior slipping
L1	22	95.5 %
L2	31	83.9 %
L3	57	93.0 %
L4	168	94.6 %
L5	74	95.9 %

$$L = L_{cls} + L_{box} + L_{mask}$$

We utilize the pre-trained ResNet50 model directly in our feature extraction network and RPN, which, as a result of its training on the ImageNet dataset, exhibits robust generalization and representation abilities. This enables the extraction of high-level features from images, a crucial aspect for both object detection and instance segmentation tasks. These networks are particularly suited for fine-tuning through transfer learning of the pre-trained models, bypassing the necessity for complete retraining. This strategy optimizes resource usage and significantly curtails training duration. To tailor the pre-trained model to our distinctive dataset it is imperative to train the object detection and segmentation networks specifically on tasks of object detection and segmentation. This necessity arises from the requirement to acquire classification and regression weights grounded in the context of lumbar instance segmentation, alongside understanding pixel-wise segmentation details.

Our dataset, exclusively composed of unique X-ray images, is partitioned with 70 % allocated for training, 20 % for testing, and 10 % reserved for validation purposes. We undertake the fine-tuning of both the feature extraction and region proposal networks, and initiate the training of the object detection and segmentation networks from scratch.

As computer vision evolves, various models refining Mask R-CNN have emerged. We consequently conducted experiments comparing these advanced models, all rooted in Mask R-CNN. Our focus: assessing their performance enhancements and practical efficacy, illuminating progress in instance segmentation methodologies.

(1) Mask Scoring R-CNN. Mask R-CNN assumes equal importance for all object masks, conflicting with the variable IoU between predicted and actual masks. Mask Scoring R-CNN tackles this by adding a branch to score masks, enhancing metric precision. This aids differentiation between accurate and inaccurate predictions, improving instance segmentation in COCO AP tests by favoring more accurate forecasts.

(2) Cascade R-CNN [28]. Cascade Mask R-CNN implements a multi-stage cascade. Each stage refines detection boxes from prior stages through dedicated regressors. By using outputs from preceding stages as inputs to subsequent ones and progressively applying stricter IoU filters, it enhances box accuracy. This cascading design mitigates overfitting and boosts detection precision, enhancing overall instance segmentation capability.

(3) Global Context-aware RoI Extractor (GRoIE) [29]. GRoIE enhances RoI extraction with a global approach, using all FPN layers for richer context compared to conventional single-layer selection. By integrating non-local features and attention, it strengthens long-range dependencies and focuses on key regions, thereby significantly upgrading the RoI extractor and model performance.

(4) Prime Sample Attention (PISA) [30]. PISA augments Mask R-CNN via the integration of Prime Sample Attention (PSA), strategically selecting informative instances based on computed relevance scores that reflect inter-sample similarity. Central to its innovation is the introduction of Prime Loss, a composite of classification and importance losses, designed to refine both segmentation accuracy and the PSA sampling strategy with a focus on pivotal “Prime Samples.” Additionally, PISA employs a concise network module to gauge mask quality, implementing IoU-HLR for ranking positive samples and Score-HLR for the hierarchical ranking of negative samples within each mini-batch, thereby further elevating instance segmentation efficacy.

After ranking, a simple Linear map is used to convert the ranking into real numbers. For class j, assuming there are a total of n_j samples ranked as $\{r_1, \dots, r_n\}$, where $0 \le r_i \le n_{max}$, convert each r_i to u_i using a linear function, as shown in the equation. Then use Exponential function to further convert the sample importance into loss weight w_i, where γ It is a degree factor that represents the priority that important samples will be given, β It is the deviation that determines the minimum sample weight.

$$u_i = \frac{n_{max} - r_i}{n_{max}}$$

$$w_i = ((1 - \beta)u_i + \beta)^{\gamma}$$

By the new weighting scheme, the classification loss L_{cls} in mask R-CNN can be rewritten as the following equation, where n and m are the number of positive and negative samples, s_i and $\widehat{s_i}$ represents the prediction score and classification objective score. In addition, in order to maintain the total loss unchanged, standardized w_i' was standardized to w_i'.

$$L_{cls} = \sum_{i=1}^{n} w_i' CE(s_i, \widehat{s_i}) + \sum_{j=1}^{m} w_j' CE(s_j, \widehat{s_j})$$

$$w_i' = w_i \frac{\sum_{i=1}^{n} CE(s_i, \widehat{s_i})}{\sum_{i=1}^{n} w_i CE(s_i, \widehat{s_i})}$$

$$w_j' = w_j \frac{\sum_{j=1}^{m} CE(s_j, \widehat{s_j})}{\sum_{j=1}^{m} w_i CE(s_j, \widehat{s_j})}$$

This ranking strategy places the positive samples with the highest IoU around each object and the negative samples with the highest score in each cluster at the top of the ranking list, and focuses the training process on these samples through a simple weighting scheme. In addition, since the quality of regression determines the importance of samples, the classifier should output higher scores for prime samples. Therefore, a classification aware regression loss is used to correlate the optimization of the two branches, in order to propagate the gradient from the regression branch to the classification branch. Using L_{carl} instead of the original regression loss $\mathscr{L}(d_i, \widehat{d_i})$, where c_i represents the prediction probability of ground truth after linear transformation.

$$L_{carl} = \sum_{i=1}^{n} c_i \mathscr{L}(d_i, \widehat{d_i})$$

Using L_{carl}, the classification branch can be supervised by regression loss. The scores of unimportant samples are greatly suppressed, while increasing attention to primary samples. Finally, this method achieved a relative improvement of 3.16 % on the bounding box mAP and 2.90 % on the instance segmentation mAP.

2.4. Image reprocessing

The instance segmentation model aforementioned accomplishes the separation of individual lumbar vertebrae from raw X-ray images. In order to facilitate the development of a subsequent lumbar spondylolisthesis diagnosis model, it is imperative to extract the coordinate points of key landmarks on the vertebrae. However, inconsistencies in facial orientations within the X-ray dataset can potentially introduce biases during feature extraction. Prior to extracting coordinates and geometric features of the vertebrae, establishing a uniform directional standard is therefore essential. Assuming that all subjects in the X-ray images face towards the right, corrective measures must be applied to images where the face is oriented to the left to align them accordingly.

Prior to correction, automated determination of the facial orientation is paramount. Given that human skeletons appear in grayscale within X-ray imagery, we initiate our approach from a grayscale perspective, employing threshold segmentation on the images. The orientation is ascertained by contrasting the areas on the left and right sides of the image. To enhance the effectiveness of threshold segmentation, histogram equalization [31] is employed to amplify image contrast and calibrate grayscale levels, facilitating clearer distinction the bone and non-bone parts of the image more distinguishable.

Subsequent to histogram equalization, the Otsu method [32] is invoked for thresholding, transforming the image into a binary format. This method relies on criteria that minimize within-class variance and maximize between-class variance for threshold determination, rendering it a favored approach for image segmentation due to its simplicity, robust adaptability, and broad applicability. Following thresholding, an image opening procedure eliminates minor internal cavities within the depicted human silhouette, culminating in the acquisition of the overall body contour. Orientation is ascertained by contrasting the sectional areas on the left and right hemispheres of the image. Fig. 2 illustrates the outcomes of histogram equalization and thresholding on the initial image, vividly demonstrating how this suite of image processing techniques readily discerns facial orientation through area comparison.

Upon ascertaining the human body's orientation, to guarantee uniformity in the feature data presented to the succeeding machine learning model, the coordinates of all anatomical landmarks within images depicting leftward-facing profiles undergo inversion. The corrected vertebral bodies are shown in the figure after orientation correction and the image after coordinate correction is shown in Fig. 3.

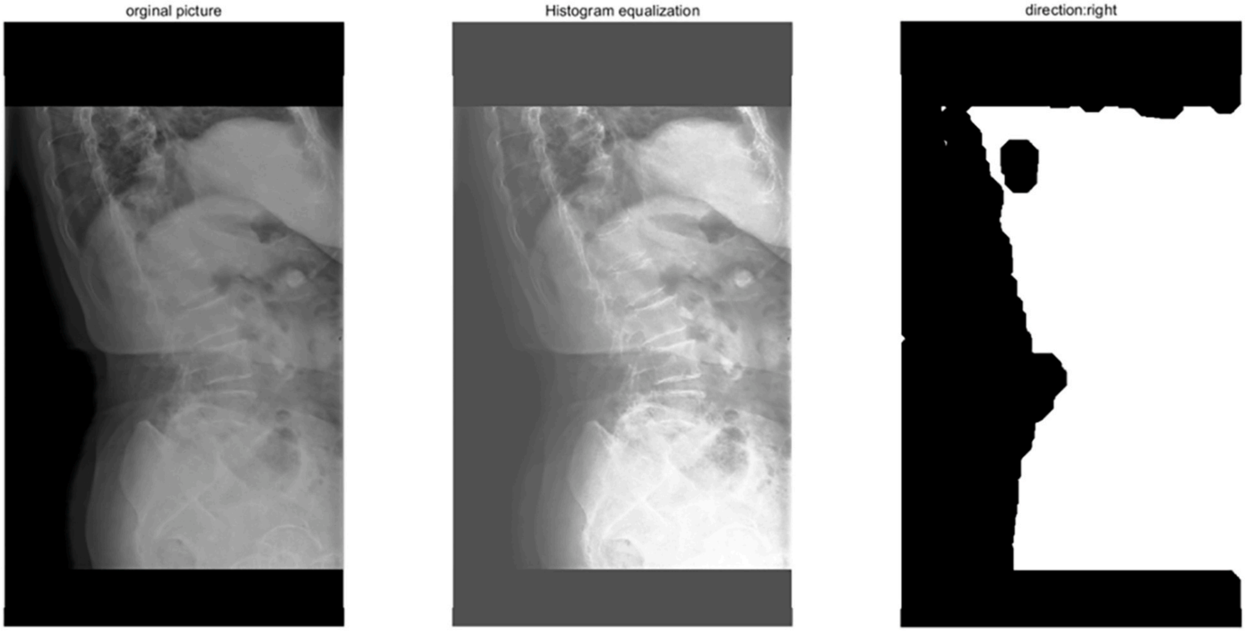

Fig. 2. Schematic diagram of the entire process of image threshold segmentation.

2.5. Development of model for diagnosis classification

After instance segmentation and coordinate correction, the diagnostic system assesses lumbar spondylolisthesis based on positional relations among adjacent vertebral landmarks. As shown in Fig. 4, the horizontal arrow in the coordinate system represents the ventral direction. Through the results of instance segmentation, we can take four points from the upper left corner, upper right corner, lower left corner, and lower right corner of each instance as our landmarks, so six landmarks (Point A, B, C, D, E, and F) are extracted from two adjacent bounding boxes, and the distance between Point G and Point D is taken as the vertex distance between the vertebrae. A straight line drawn from Point A to Point B was expressed as AB, and the distance between Point A and Point B was expressed as L_{AB}. Based on the 4 extracted lines (AB, BC, DE, DF) between the landmarks, the length and slope (with reference to the horizontal direction) of each line are calculated. Simultaneously, grounded in medical precedent, the "Slip Ratio" (the same as **P-Grade** [21]) is deduced by scrutinizing the correlation between superior and inferior vertebral landmarks. Meyerding's and Taillard's method [7,33] was referred to measure the Slip Ratio, the value was calculated as L_{DG}/L_{DE} (Fig. 4). This ratio represents the coherence of the vertebrae from top to bottom. Since we have three X-rays of different views for each vertebra, along with the serial number vertebra, there are a total of 28 dimensions of features.

Since the directly extracted features are relatively intuitive and low-level, and the actual lumbar spondylolisthesis is determined by the relative position relationship between these lines, feature crossing is used to explore deeper features for model training. The specific implementation of feature crossing is to directly multiply or divide the normalized features from different domains. During model training, the model automatically retains the second-order features that have a significant impact on the results. To overcome the imbalance in the distribution of original samples, the SMOTE oversampling method [34] is used to enhance the samples of spondylolisthesis. Random oversampling algorithms are prone to overfitting, which can lead to models that are too specific and not general enough. SMOTE oversampling changes the data distribution of imbalanced datasets by adding generated samples of the minority class, which is a popular method for improving the performance of classification models on imbalanced data. SMOTE synthesizes new samples by linear interpolation between two minority class samples, effectively alleviating the overfitting caused by random oversampling.

From the outcomes of instance segmentation, landmarks are extracted, followed by the derivation of 28-dimensional features from the intervertebral region. These features undergo normalization and crossing to enrich the feature space, facilitating the training of machine learning models. Considering the modest dimensionality of samples and moderate dataset size, ensemble tree models prove more advantageous over deep learning models in this context. Consequently, we use the Light Gradient Boosting Machine (LGBM) [35] model as the classifier. Within the realm of ensemble tree models, LGBM and XGBoost[35] have garnered prominence for their exemplary efficiency and versatility. Owing to their shared foundation in decision tree-based gradient boosting, LGBM and XGBoost each command a distinct, influential role as elite classifiers in the machine learning sphere, with remarkable performance especially in classifications.

2.6. Evaluation and interpretability analysis of models

Mean Average Precision(mAP), widely used in COCO format datasets [36], was applied to evaluate the performance of the models for instance segmentation. It included bounding box mAP(Bbox mAP) and segmentation mAP(Segm mAP), and mAP at different IoU (mAP50 and mAP75).

To evaluate the performance of the models for diagnosis of lumbar spondylolisthesis, the accuracy, recall, precision, F1-score of the model were evaluated and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve was calculated, too.

To further understand which feature contribute more to the results in the model, SHAP values were used to interpret the LGBM model and help us understand the impact of each feature on the predicted outcome. The Shapley value from game theory is used to allocate the contribution of each feature to the predicted outcome. For each predicted sample, the model generates a predicted value, and the SHAP value [37] is the value assigned to each feature in that sample. The idea behind SHAP values is to allocate total revenue

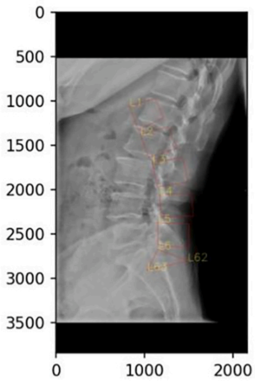

Fig. 3. Example of an image with coordinate correction based on facial orientation.

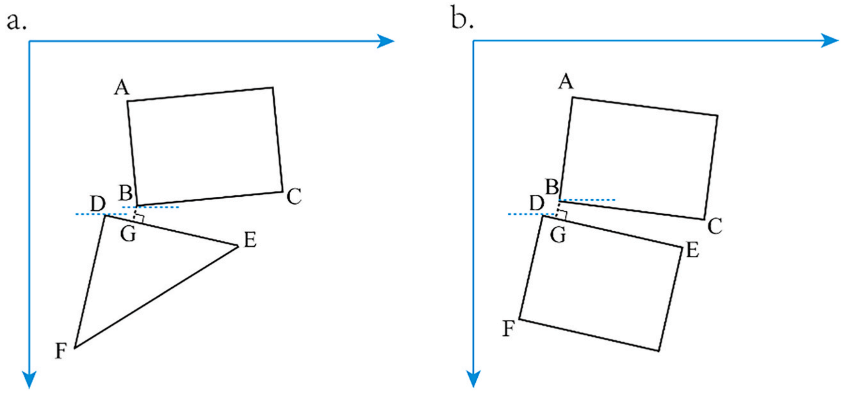

Fig. 4. Schematic diagram of the features, including AB, BC, DE, DF, and the Slip Ratio equals L_{DG}/L_{DE}. a. Schematic diagram of L5/S1. b. Schematic diagram of other segments.

by calculating the expected marginal contribution of each feature to all possible coalitions.

Model construction, evaluation and analysis of the validation measures were performed using Python 3.8.8, and the instance segmentation model is trained and predicted using the mmdetection framework [38] version 2.0.0, with one NVIDIA GeForce RTX 3060 GPU device and pytorch 1.8.0. Machine learning models (including LGBM and XGBoost etc.) are trained and predicted by the pycaret 2.3.5 framework [39].

2.7. Code availability

Our code has been made publicly available at https://github.com/THUzyt21/Diagnosis-of-lumbar-spondylolisthesis-with-LGBM. Details regarding feature processing and the diagnostic model can be referenced there.

3. Results

3.1. Performance of models

Regarding improvements to Mask R-CNN in various strategies, for our dataset, the most effective is the Prime Sample Attention strategy. It is an importance-based sample reweighting method that can increase the model's attention to important samples, thereby improving the mAP of the instance segmentation model. In tests on 190 images in the test set, the bbox and segm mAP of the models are shown in Table 2 and the instance segmentation results on the test set are shown in Fig. 5.

After extracting the image features using the feature extraction method mentioned earlier, we trained deep learning models and post-processing tree models using 65 % of the data as the training set, and performed model predictions on the remaining 35 % of the data. The best performing model was Light Gradient Boosting Machine, with an accuracy of 96.60 %, an AUC of 0.9843, a F1-score of 0.9020. We compared these machine learning models, and the relevant results are shown in Table 3. In our model configuration, key parameters include a gradient boosting decision tree (boosting_type = 'gbdt') with a learning rate of 0.1, 100 estimators, a maximum of 31 leaves per tree, and a subsample ratio of 1.0 for both data and features, ensuring comprehensive utilization of the dataset during training with controlled complexity for balanced predictive performance.

Fig. 6 illustrates the performance of the automated diagnostic system, employing the optimal instance segmentation and classification models, where an accuracy of 96.60 % and a precision of 90.20 % were attained, showcasing its efficacy.

To benchmark our machine learning models against prevailing methodologies in diagnosing lumbar spondylolisthesis via X-rays, we adopted and implemented the process outlined by Trinh et al. [21], involving manual extraction of lumbar geometric features and subsequent diagnostic inference through rule-based systems, adhering to their defined diagnostic parameters for the lumbar spine. The comparison of our method's results with theirs is presented in Table 4.

Table 2
Comparison of metrics for Mask R-CNN and its various improved models in different directions.

model	Bbox mAP	Bbox mAP50	Bbox mAP75	Segm mAP	Segm mAP50	Segm mAP75
mask R-CNN	0.7468	0.9831	0.8856	0.7242	0.9725	0.8708
cascade mask R-CNN	0.7758	0.9832	0.9231	0.7354	0.9724	0.8769
mask R-CNN + GRoIE	0.7461	0.9836	0.9044	0.7376	0.9724	0.8804
mask scoring R-CNN	0.7561	0.9827	0.9071	0.7405	0.9540	0.8732
mask R-CNN + prime sample attention	0.7704	0.9852	0.9179	0.7452	0.9741	0.8957

Bbox mAP: bounding box mean average precision, Segm mAP: Segmentation mean average precision.

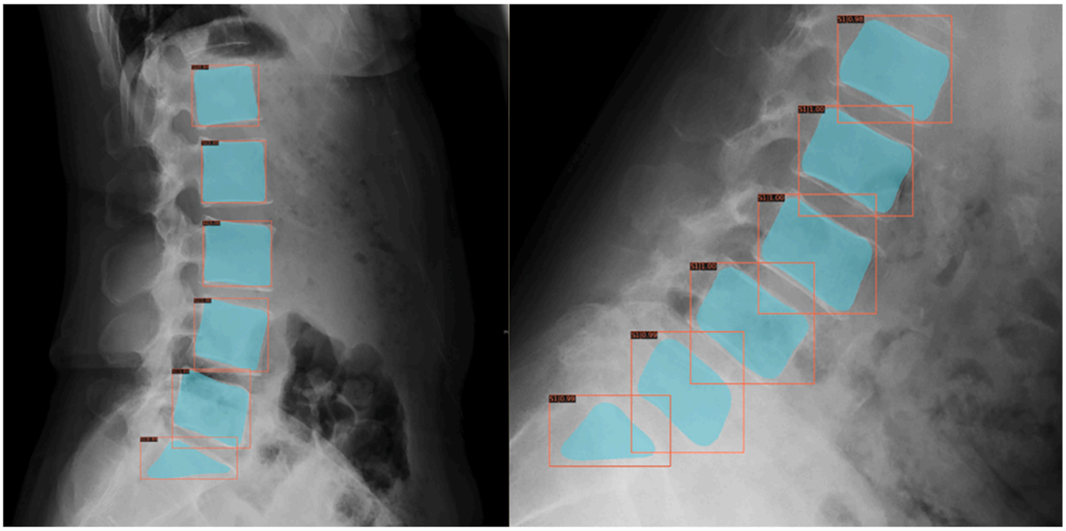

Fig. 5. Schematic diagram of lumbar spine X-ray angiography results after instance segmentation model.

Table 3
Comparison of the effectiveness of various machine learning models in diagnosing lumbar vertebral slip.

Model	Accuracy	AUC	Recall	Precision	F1-score
LGBM	0.9660	0.9843	0.9020	0.9020	0.9020
GBC	0.9558	0.9739	0.9020	0.8519	0.8762
RF	0.9558	0.9839	0.9216	0.8393	0.8785
ADA	0.9490	0.9726	0.9020	0.8214	0.8598
ET	0.9592	0.9836	0.9020	0.8679	0.8846
XGBoost	0.9354	0.9697	0.9020	0.7667	0.7933
DT	0.9116	0.9275	0.8627	0.6984	0.9116
KNN	0.8878	0.8910	0.8235	0.6364	0.7179
SVM	0.7789	0.8353	0.9216	0.4352	0.5912

LGBM: Light Gradient Boosting Machine, GBC: Gradient Boosting Classifier, RF: Random Forest Classifier, ADA: Ada Boost Classifier, ET: Extra Trees Classifier, XGBoost: Extreme Gradient Boosting, DT: Decision Tree Classifier, KNN: K Neighbors Classifier, SVM: Support Vector Machine.

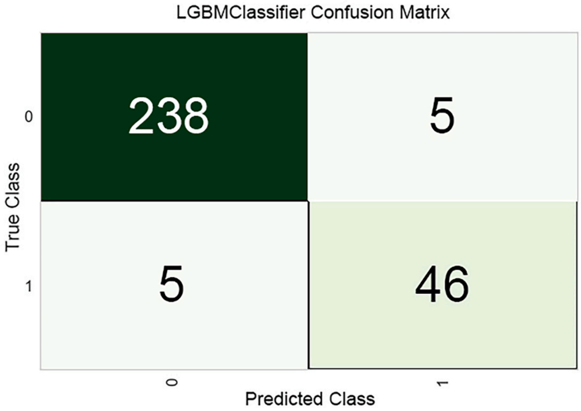

Fig. 6. Confusion Matrices of automatic diagnosis by LGBM Classifier.

An interpretive examination of the trained LGBM model reveals that the paramount features, as depicted in Fig. 7, are the vertebral sequence number and slippage ratio, echoing these factors' significance from diverse viewpoints - an alignment with established medical understanding.

Table 4
Comparison of diagnostic performance between our machine learning approach and Trinh's rule-based method.

Metrics	P-grade(k1 = 10)	PSD(K2 = 37)	P-Grade + PSD	DS(K3 = 0.14)	Ours
acc	0.8880	0.8749	0.8856	0.7104	0.9660
precision	0.7692	0.8842	0.7384	0.2400	0.9020
recall	0.6742	0.4719	0.7135	0.1685	0.9020
F1-score	0.7186	0.6154	0.7257	0.1980	0.9020

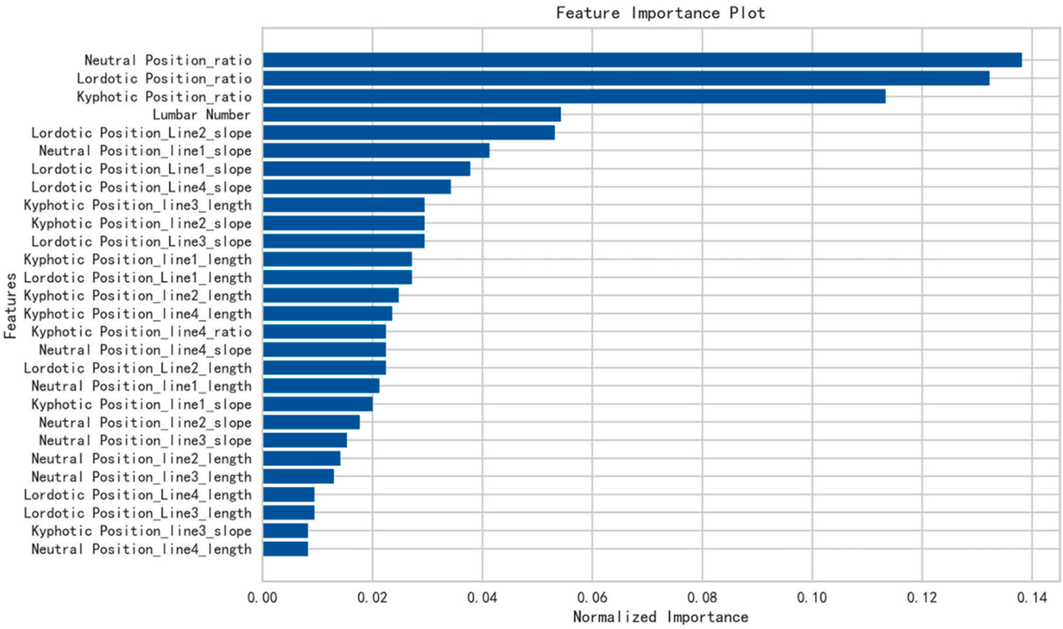

Fig. 7. Schematic diagram of feature selection results.

4. Discussion

In this study, we harnessed the power of Mask R-CNN and used prime sample attention to precisely delineate the position of the vertebral body. Based on the outputs of instance segmentation, we meticulously pinpointed four key coordinates which served as the foundation for extracting pertinent features related to vertebral body positioning. These extracted features provide valuable insights into the spatial relationships of the lumbar vertebrae. Feature data were subsequently employed to train a diverse array of machine learning models. Following a rigorous comparative analysis, LGBM proved to be the most adept model for classifying spondylolisthesis.

Different from previous studies that just used a single model and had a single function of diagnosis whether a patient had lumbar spondylolisthesis or not, the diagnostic system constructed in our study extends this capability. The system not only shows the region of target segment but also output the serial number of the spondylolisthesis, which means it has the function of drawing a complete diagnostic conclusion of the disease. The function is based on a series of two machine learning models, namely, an instance segmentation model for identifying the position of vertebral bodies and a tree-based model for diagnosis. After performing the instance segmentation to obtain the coordinates of landmarks of different vertebral bodies, we modified the Taillard's method [33], and introduced a one-dimensional coordinate axis with direction when measure the displacement. By using a single feature of the segmentation model, the results may be affected by artifacts and the diagnostic performance would not be robust. Therefore, we extracted more features in all 3 positions images from the instance segmentation model. After feature normalization, crossing and combining, these features were acquired to train new classification models, and the model with the best performance was selected after comparison.

In the trained LGBM model, feature importance analysis revealed that the most critical feature is the slip ratio measured on different lateral X-ray views, with the slip ratio in three positions contributing 38.4 % to the feature importance (Fig. 7). The importance of the slip ratio aligns with clinical cognition. By definition, slip ratio effectively reflects the occurrence of vertebral slippage. Meyerding [7] used it to assess the severity of spondylolisthesis in patients, and similarly, Trinh et al. [21] utilized it as a feature for the automated diagnosis of lumbar spondylolisthesis. The lumbar segment number is also an important feature in our model, which correlates with the varying incidence rates of spondylolisthesis across different lumbar segments. Previous studies have demonstrated significant differences in the probability of spondylolisthesis occurrence among different segments [40,41]. In clinical practice, surgeons pay particular attention to segments with higher incidence rates to reduce missed diagnoses of spondylolisthesis.

Therefore, this model aligns well with clinical cognition. Among other important features of the LGBM model, the first-order and second-order features related to the Slip Ratio also account for the vast majority. Therefore, the classification model can well explain the definition, degree, segmental difference and posture difference of spondylolisthesis, which is consistent with clinical facts and has strong practical value.

To the best of our knowledge, there are scarce studies that have concatenated instance segmentation models with classification models for the automatic diagnosis of lumbar spondylolisthesis, explicitly outlining the affected vertebral segment numbers. Several researches on automatic diagnosis of lumbar spondylolisthesis by analyzing imaging data have published. Hu et al. [42] proposed a new neural network structure named Swin-PGNet, which was trained using annotated X-ray images, so that it can automatically locate the landmarks of the lumbar vertebral body, and measure the L4 lumbar sliding distance through the predicted landmarks. The diagnostic accuracy of the model for spondylolisthesis was 71.3 %. The model can automatically obtain the coordinates of the corresponding landmarks by learning the annotated images, and then use these coordinates to calculate the distance of displacement directly, which is similar to the measure method of the first model used in our study. Although the accuracy rate of using a single model is higher than physician in the same study (71.3 % VS 70.7 %), but compared with other machine learning models, the performance is not so good. Another limitation is the model in this study only analyzed a single segment of L4, which may not so accurate when applied to diagnosis of spondylolisthesis in other segments. Lehnen et al. [43] applied deep neural network to automatic diagnosis of lumbar degenerative diseases including spondylolisthesis. In this study, U-Net-based CNN was used in their study to segment MRI images. Segmentation results was applied to identify different diseases according to their geometric relationship. For spondylolisthesis diagnosing, they employed Meyerding's classification, which is primarily intended for grading the severity of spondylolisthesis after diagnosis. In fact, Meyerding's classification does not include the definition that a slip ratio greater than 0 can diagnose spondylolisthesis. This misuse might be a major reason for the substantial disagreement between the model and experts in diagnosing spondylolisthesis, resulting in a positive predictive value of only 13.11 %. Zhao et al. [44] also trained a model for grading lumbar spondylolisthesis using MRI images. However, this model similarly cannot perform the diagnostic task for lumbar spondylolisthesis. Additionally, MRI-based models, due to their high cost and time-consuming nature, are not the primary imaging choice for diagnosing lumbar spondylolisthesis in clinical practice. Consequently, they may not be widely used in outpatient care or in large-scale population disease screening.

In contrast to these approaches, Trinh et al. [21] employed image segmentation to extract key points of the lumbar vertebrae and manually constructed features for diagnosing spondylolisthesis, thereby enabling the automatic identification of the specific affected segment. Nonetheless, their limitation resides in the insufficient automation of manually computed features, with certain pivotal diagnostic parameters (akin to K1, K2, K3 in this study) requiring manual specification. Moreover, automated feature learning excels at uncovering latent patterns and intricate relationships within data that might elude straightforward human recognition or construction. Furthermore, features automatically extracted by machine learning models tend to generalize better across different datasets, whereas manually crafted features can be overly tailored to specific tasks or datasets, potentially compromising generalization performance when applied to novel contexts. Indeed, their method displayed minor discrepancies in performance when applied to our dataset compared to their own. Lastly, while manually defined features hold an advantage in terms of interpretability, being grounded in explicit medical expertise, our machine learning model leverages SHAP values for feature explanation, thereby mitigating, to a certain extent, the traditional disadvantage of poor explainability associated with machine learning methodologies.

Compared with these methods, this study has engineered an integrated diagnostic system that combines an instance segmentation model, medical feature extraction methodologies, and a classifier, thereby harnessing the strengths of each component to significantly enhance diagnostic accuracy in comparison to the diagnostic pipeline employed by Trinh et al. [21] When combined with the efficiency advantage of artificial intelligence, it significantly enhances the accuracy and efficiency of clinical assessment and scientific research pertaining to lumbar spondylolisthesis.

5. Limitations

This study has certain limitations that warrant improvement. Firstly, the system was trained and validated using a single-center, retrospective dataset, which could introduce selection bias. Secondly, the standardization and reprocessing of X-ray images can be time-consuming, necessitating more streamlined approaches in the future. Additionally, our system is currently unable to diagnose the retrolisthesis due to limited data availability for model training and testing within our dataset.

6. Conclusion

In conclusion, this study presents a robust system for the diagnosis of lumbar spondylolisthesis, capable of not only accurately detecting the presence of spondylolisthesis but also precisely identifying the segmental sequence number of the slippage. Moreover, our approach introduces a novel method for automated quantitation of spondylolisthesis. The pipeline developed in this study may also hold promise for identifying other spinal diseases based on contour detection, potentially leading to the development of more versatile diagnostic models.

Ethics declarations

Approval for the study was obtained from the Peking University Third Hospital Institutional Review Board in March 3rd, 2022 (S2022290). Informed consent was waived due to the retrospective nature of the investigation.

Funding/Support

This study was funded by the National Natural Science Foundation of China (No. 82272577) and Key Clinical Projects of Peking University Third Hospital (BYSYZD2021024). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data availability statement

To protect patient privacy and confidentiality, clinical data will not be publicly available. However, these data can be requested from the corresponding author upon reasonable demand.

CRediT authorship contribution statement

Shanshan Liu: Writing – review & editing, Formal analysis, Data curation, Conceptualization. **Chenyi Guo:** Writing – review & editing, Formal analysis, Data curation, Conceptualization. **Yuting Zhao:** Writing – original draft, Visualization, Investigation, Data curation. **Cheng Zhang:** Writing – original draft, Visualization, Investigation, Formal analysis. **Lihao Yue:** Visualization, Investigation. **Ruijie Yao:** Visualization, Software. **Qifeng Lan:** Visualization, Investigation. **Xingyu Zhou:** Investigation. **Bo Zhao:** Investigation. **Ji Wu:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization. **Weishi Li:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization. **Nanfang Xu:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

All authors declare no financial or non-financial competing interests.

References

[1] Z. Lan, J. Yan, Y. Yang, Q. Xu, Q. Jin, A review of the main classifications of lumbar spondylolisthesis, World neurosurgery 171 (2023) 94–102.

[2] N.V. Mohile, A.S. Kuczmarski, D. Lee, C. Warburton, K. Rakoczy, A.J. Butler, Spondylolysis and isthmic spondylolisthesis: a guide to diagnosis and management, J. Am. Board Fam. Med. : JABFM 35 (2022) 1204–1216.

[3] I.M. Austevoll, E. Hermansen, M.W. Fagerland, K. Storheim, J.I. Brox, T. Solberg, F. Rekeland, E. Franssen, C. Weber, H. Brisby, O. Grundnes, K.R.H. Algaard, T. Böker, H. Banitalebi, K. Indrekvam, C. Hellum, Decompression with or without fusion in degenerative lumbar spondylolisthesis, N. Engl. J. Med. 385 (2021) 526–538.

[4] F.L. Wei, C.P. Zhou, Q.Y. Gao, M.R. Du, H.R. Gao, K.L. Zhu, T. Li, J.X. Qian, X.D. Yan, Decompression alone or decompression and fusion in degenerative lumbar spondylolisthesis, EClinicalMedicine 51 (2022) 101559.

[5] W.C. Watters 3rd, C.M. Bono, T.J. Gilbert, D.S. Kreiner, D.J. Mazanec, W.O. Shaffer, J. Baisden, J.E. Easa, R. Fernand, G. Ghiselli, M.H. Heggeness, R.C. Mendel, C. O'Neill, C.A. Reitman, D.K. Resnick, J.T. Summers, R.B. Timmons, J.F. Toton, An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spondylolisthesis, Spine J. 9 (2009) 609–614.

[6] P.G. Matz, R.J. Meagher, T. Lamer, W.L. Tontz Jr., T.M. Annaswamy, R.C. Cassidy, C.H. Cho, P. Dougherty, J.E. Easa, D.E. Enix, B.A. Gunnoe, J. Jallo, T. D. Julien, M.B. Maserati, R.C. Nucci, J.E. O'Toole, K. Rosolowski, J.N. Sembrano, A.T. Villavicencio, J.P. Witt, Guideline summary review: an evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spondylolisthesis, Spine J. 16 (2016) 439–448.

[7] H. Meyerding, Spondylolisthesis, Surg Gynecol Obstet, vol. 54, 1932, pp. 371–377.

[8] L.L. Wiltse, P.H. Newman, I. Macnab, Classification of spondylolisis and spondylolisthesis, Clin. Orthop. Relat. Res. (1976) 23–29.

[9] S. Butt, A. Saifuddin, The imaging of lumbar spondylolisthesis, Clin. Radiol. 60 (2005) 533–546.

[10] M. Fraiwan, Z. Audat, L. Fraiwan, T. Manasreh, Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images, PLoS One 17 (2022) e0267851.

[11] B.H. Cho, D. Kaji, Z.B. Cheung, I.B. Ye, R. Tang, A. Ahn, O. Carrillo, J.T. Schwartz, A.A. Valliani, E.K. Oermann, V. Arvind, D. Ranti, L. Sun, J.S. Kim, S.K. Cho, Automated measurement of lumbar lordosis on radiographs using machine learning and computer vision, Global Spine J. 10 (2020) 611–618.

[12] M.H. Horng, C.P. Kuok, M.J. Fu, C.J. Lin, Y.N. Sun, Cobb angle measurement of spine from X-ray images using convolutional neural network, Comput. Math. Methods Med. 2019 (2019) 6357171.

[13] L. Zou, L. Guo, R. Zhang, L. Ni, Z. Chen, X. He, J. Wang, VLTENet: a deep-learning-based vertebra localization and tilt estimation network for automatic cobb angle estimation, IEEE journal of biomedical and health informatics 27 (2023) 3002–3013.

[14] R. Zhang, Y. Hu, K. Zhang, G. Lan, L. Peng, Y. Zhu, W. Qian, Y. Yao, VDVM: an automatic vertebrae detection and vertebral segment matching framework for C-arm X-ray image identification, J. X Ray Sci. Technol. 31 (2023) 935–949.

[15] H. Wang, T. Zhang, K.M. Cheung, G.K. Shea, Application of deep learning upon spinal radiographs to predict progression in adolescent idiopathic scoliosis at first clinic visit, EClinicalMedicine 42 (2021) 101220.

[16] Y. Shin, K. Han, Y.H. Lee, Temporal trends in cervical spine curvature of south Korean adults assessed by deep learning system segmentation, 2006-2018, JAMA Netw. Open 3 (2020).

[17] Y. Okita, T. Hirano, B. Wang, Y. Nakashima, S. Minoda, H. Nagahara, A. Kumanogoh, Automatic evaluation of atlantoaxial subluxation in rheumatoid arthritis by a deep learning model, Arthritis Res. Ther. 25 (2023) 181.

[18] X. Chen, Q. Deng, Q. Wang, X. Liu, L. Chen, J. Liu, S. Li, M. Wang, G. Cao, Image quality control in lumbar spine radiography using enhanced U-net neural networks, Front. Public Health 10 (2022) 891766.

[19] S. Al Arif, K. Knapp, G. Slabaugh, Fully automatic cervical vertebrae segmentation framework for X-ray images, Comput. Methods Progr. Biomed. 157 (2018) 95–111.

[20] J. Zhang, H. Lin, H. Wang, M. Xue, Y. Fang, S. Liu, T. Huo, H. Zhou, J. Yang, Y. Xie, M. Xie, L. Cheng, L. Lu, P. Liu, Z. Ye, Deep learning system assisted detection and localization of lumbar spondylolisthesis, Front. Bioeng. Biotechnol. 11 (2023) 1194009.

[21] G.M. Trinh, H.C. Shao, K.L. Hsieh, C.Y. Lee, H.W. Liu, C.W. Lai, S.Y. Chou, P.I. Tsai, K.J. Chen, F.C. Chang, M.H. Wu, T.J. Huang, Detection of lumbar spondylolisthesis from X-ray images using deep learning network, J. Clin. Med. 11 (2022).

[22] W.M. Durand, R. Lafage, D.K. Hamilton, P.G. Passias, H.J. Kim, T. Protopsaltis, V. Lafage, J.S. Smith, C. Shaffrey, M. Gupta, M.P. Kelly, E.O. Klineberg, F. Schwab, J.L. Gum, G. Mundis, R. Eastlack, K. Kebaish, A. Soroceanu, R.A. Hostin, D. Burton, S. Bess, C. Ames, R.A. Hart, A.H. Daniels, Artificial intelligence clustering of adult spinal deformity sagittal plane morphology predicts surgical characteristics, alignment, and outcomes, Eur. Spine J. 30 (2021) 2157–2166.

[23] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 386–397.

[24] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149.

[25] S. Zhao, B. Chen, H. Chang, B. Chen, S. Li, Reasoning discriminative dictionary-embedded network for fully automatic vertebrae tumor diagnosis, Med. Image Anal. 79 (2022) 102456.

[26] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 640–651.

[27] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764–773.

[28] Z. Cai, N. Vasconcelos, Cascade R-CNN: high quality object detection and instance segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 1483–1498.

[29] L. Rossi, A. Karimi, A. Prati, A novel region of interest extraction layer for instance segmentation, in: International Conference on Pattern Recognition (ICPR), 2021, pp. 2203–2209.

[30] Y. Cao, K. Chen, C.C. Loy, D. Lin, Prime sample attention in object detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11580–11588.

[31] L. Lu, Y. Zhou, K. Panetta, S.S. Agaian, Comparative study of histogram equalization algorithms for image enhancement, Mobile Multimedia/Image Processing, Security, and Applications 2010 7708 (2010) 770811.

[32] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1979) 62–66.

[33] W. Taillard, [Spondylolisthesis in children and adolescents], Acta Orthop. Scand. 24 (1954) 115–144.

[34] A. Fernandez, S. Garcia, N.V. Chawla, F. Herrera, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, J. Artif. Intell. Res. 61 (2018) 863–905.

[35] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, 2017, pp. 3149–3157.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.

[37] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, ArXiv, 2017 abs/1705.07874.

[38] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C.C. Loy, D. Lin, MMDetection: Open MMLab Detection Toolbox and Benchmark, ArXiv, 2019 abs/1906.07155.

[39] M. Ali, PyCaret: an Open Source, Low-Code Machine Learning Library in Python, 2020.

[40] L. Kalichman, D.H. Kim, L. Li, A. Guermazi, V. Berkin, D.J. Hunter, Spondylolysis and spondylolisthesis: prevalence and association with low back pain in the adult community-based population, Spine 34 (2009) 199–205.

[41] Y. Aoki, H. Takahashi, A. Nakajima, G. Kubota, A. Watanabe, T. Nakajima, Y. Eguchi, S. Orita, H. Fukuchi, N. Yanagawa, K. Nakagawa, S. Ohtori, Prevalence of lumbar spondylolysis and spondylolisthesis in patients with degenerative spinal disease, Sci. Rep. 10 (2020) 6739.

[42] H. Hu, X. Wang, H. Yang, J. Zhang, K. Li, J. Zeng, [Development and validation of an automatic diagnostic tool for lumbar stability based on deep learning], Zhongguo Xiu Fu Chong Jian Wai Ke Za Zhi 37 (2023) 81–90.

[43] N.C. Lehnen, R. Haase, J. Faber, T. Ruber, H. Vatter, A. Radbruch, F.C. Schmeel, Detection of degenerative changes on mr images of the lumbar spine with a convolutional neural network: a feasibility study, Diagnostics 11 (2021).

[44] S. Zhao, X. Wu, B. Chen, S. Li, Automatic spondylolisthesis grading from MRIs across modalities using faster adversarial recognition network, Med. Image Anal. 58 (2019) 101533.