mathematics

MDPI

Article

Deep Learning-Based Intelligent Diagnosis of Lumbar Diseases with Multi-Angle View of Intervertebral Disc

Kaisi (Kathy) Chen [1,2], Lei Zheng [3], Honghao Zhao [1,*] and Zihang Wang [1]

[1] Department of Decision Sciences, School of Business, Macau University of Science and Technology, Macao 999078, China; ccchenkaisi@gmail.com (K.C.); 2220006702@student.must.edu.mo (Z.W.)
[2] Department of Mathematics and Statistics, University of Canterbury, Christchurch 8041, New Zealand; kaisi.chen@pg.canterbury.ac.nz (K.C.)
[3] Department of Management, School of Business, Macau University of Science and Technology, Macao 999078, China; zhenglei@must.edu.mo
* Correspondence: hhzhao@must.edu.mo

Abstract: The diagnosis of degenerative lumbar spine disease mainly relies on clinical manifestations and imaging examinations. However, the clinical manifestations are sometimes not obvious, and the diagnosis of medical imaging is usually time-consuming and highly relies on the doctor's personal experiences. Therefore, a smart diagnostic technology that can assist doctors in manual diagnosis has become particularly urgent. Taking advantage of the development of artificial intelligence, a series of solutions have been proposed for the diagnosis of spinal diseases by using deep learning methods. The proposed methods produce appealing results, but the majority of these approaches are based on sagittal and axial images separately, which limits the capability of different deep learning methods due to the insufficient use of data. In this article, we propose a two-stage classification process that fully utilizes image data. In particular, in the first stage, we used the Mask RCNN model to identify the lumbar spine in the spine image, locate the position of the vertebra and disc, and complete rough classification. In the fine classification stage, a multi-angle view of the intervertebral disc is generated by splicing the sagittal and axial slices of the intervertebral disc up and down based on the key position identified in the first stage, which provides more pieces of information to the deep learning methods for classification. The experimental results reveal substantial performance enhancements with the synthesized multi-angle view, achieving an F1 score of 96.67%. This represents a performance increase of approximately 15% over the sagittal images at 84.48% and nearly 14% over the axial images at 83.15%. This indicates that the proposed paradigm is feasible and more effective in identifying spinal-related degenerative diseases through medical images.

Keywords: degenerative lumbar spine disease; multi-angle view of disc; mask RCNN model; two-staged classification

MSC: 68U10

Citation: Chen, K.; Zheng, L.; Zhao, H.; Wang, Z. Deep Learning-Based Intelligent Diagnosis of Lumbar Diseases with Multi-Angle View of Intervertebral Disc. *Mathematics* **2024**, *12*, 2062. https://doi.org/10.3390/math12132062

Academic Editors: Filippo Geraci, Marco Fornili and Demetrio Labate

Received: 22 April 2024
Revised: 26 June 2024
Accepted: 28 June 2024
Published: 1 July 2024

1. Introduction

With the global modernization of science and technology, prolonged sitting and standing are becoming increasingly common, resulting in an increased incidence of diseases such as low back pain, lumbar disc herniation, and vertebral bone spur degeneration. Lumbar diseases can occur at all ages, mainly in elderly individuals, but in recent years, an increasing number of young people also suffer from lumbar diseases. The prevalence of spinal diseases is not only high but also has serious consequences. Low back pain caused by lumbar disc herniation is the most common disease, with a total incidence rate of 15–30% in Western countries and as high as 18% in China. Ref. [1], based on the 1990–2016 disease burden report in China and various provinces, quantified the prevalence and years of disability caused by low back pain in China. He proposed that low back

pain is the world's most common cause of disability in people and is also the second largest cause of disability burden among Chinese people. Therefore, considerable attention to LBP is needed, especially for the female population. According to the U.S. National Institutes of Health (NIH), the incidence of degenerative diseases of the spine is as high as 5.7% worldwide. Outpatient epidemiological surveys show that the incidence of spinal diseases is second only to influenza, and approximately 76% of outpatients experience symptoms of low back pain. In March 2018, Refs. [2–4] published three reports in the leading international medical journal The Lancet, calling attention to the global problem of lower-back pain and urgent action. Therefore, people must pay attention to the prevention and treatment of lumbar disease. At the same time, the diagnosis of spinal disease is time-consuming and subjective—radiologists look at images of patients in different frames to determine what is causing the patient's clinical presentation. Sometimes doctors come to different conclusions about the same disease. The diagnostic process for spinal diseases is so cumbersome that the implementation of intelligent diagnostic systems to assist doctors is urgent. The difference between normal and suspected patients can be determined before the radiologist can examine them, which will greatly reduce the workload of doctors and the cost of testing. It can allow people with slight discomfort to check their bodies in advance without considering money, to prevent the occurrence of disease in advance.

Medical imaging research is one of the solutions to these problems, and magnetic resonance imaging (MRI) is generally used as source data for clinical decision-making. Clinical diagnoses made by doctors are usually based on their personal experience, and intelligent diagnostic methods or systems provide effective support for the doctors. With the rapid development of artificial intelligence, machine learning techniques have become popular and widely accepted by doctors in the field of supporting clinical diagnosis. Ref. [5] proposed a new fuzzy method called interval type-2 fuzzy regression (IT2FR) to diagnose schizophrenia (SZ) and attention deficit hyperactivity disorder (ADHD). Ref. [6] proposed a new deep learning method with robust features learned from a generative adversarial network (GAN) on different datasets of MR images for the purpose of tumor classification. With the success of previous research, an increasing number of clinical diagnostic applications have sought support with the proper use of machine learning methods. Diagnostic studies of spinal diseases also fall into this group. Ref. [7] investigated the effectiveness of three dimensionality reduction techniques and three feature-selection techniques for classifying lumbar spine MRI. Ref. [8] proposed a classifier based on a deep convolutional neural network and considered extensive hyperparameter optimization to improve the performance of intervertebral disc classification. Ref. [9] considered segmentation of the region of interest (ROI) in the disc area to improve the classification performance using a convolutional neural network.

The above-mentioned studies have produced promising results for the diagnosis of spinal diseases by analyzing magnetic resonance imaging with modern deep-learning approaches. However, they still have some common drawbacks. First, the majority of research in the literature has focused mainly on the diagnosis of intervertebral discs, and few scholars classify both discs and vertebrae. More importantly, their proposed methods are limited by the insufficient use of MRI data. Normally, in the literature, the classification results generated by different deep learning approaches are based on the sagittal and axial view planes independently, which fails to consider the correlations between images of different planes. For example, Ref. [10] proposed a novel regional feature recalibration network (RFRecNet) is proposed to achieve accurate and reliable lumbar intervertebral disc degeneration grading. Ref. [11] proposed a new way for lumbar disc location. These latest studies still focus on accuracy improvement given a single data source. However, the fusion of multi-angle information can provide a deeper understanding of the nature of the target, which is essentially helpful in the classification process when using different deep-learning approaches. Therefore, this paper proposes a two-stage process to classify the lumbar spine images through deep learning approaches. In particular, in the first stage, the Mask RCNN model was used to identify the lumbar spine in the spine image, locate

the positions of the vertebra and disc, and complete their rough classification. In the fine classification stage, a synthesized multi-angle view of the intervertebral disc is generated by splicing the sagittal and axial slices of the intervertebral disc up and down based on the key positions identified in the first stage. Finally, the VGG, ResNet, and MobileNet models were introduced for the fine classification of spinal diseases. The entire process allows deep learning approaches to fully use the information of the MRI image data.

2. Related Works

In the existing clinical diagnosis studies of spinal diseases, relevant studies on the application of machine learning technology can be divided into the following three parts: (1) The segmentation of the spine, including the segmentation of the vertebrae and the intervertebral disc; (2) The localization and labeling of the vertebrae; (3) Clinical auxiliary diagnosis of spinal diseases, including spinal deformity, spinal degenerative diseases, outcome prediction, and clinical gait analysis. In the following, a literature review is delivered based on the three aspects mentioned above.

2.1. Research on Medical Image Segmentation

The segmentation of medical images plays an important role in pathological research on different parts of the human body. Before performing pathological analysis, doctors need to determine the location of the diseased area and analyze whether the characteristics of the diseased area correspond to the characteristics of the previous medical records to determine the type of disease. Similarly, for a computer, it is necessary to determine the region of interest before the extraction features of this region can be judged and analyzed. Therefore, the accuracy of judgment largely depends on the effectiveness of segmentation.

Ref. [12] divided medical image segmentation into two categories: semi-automatic segmentation and automated segmentation. He proposed that the future segmentation of medical images should be based on the premise of reducing manual intervention and improving the computational efficiency and accuracy of the segmentation algorithm. In recent years, the segmentation of spine lesions has progressively switched from the original manual segmentation to fully automatic deep learning-based segmentation. Spine segmentation was originally based on the shape model. Ref. [13] employed an improved machine learning algorithm, AdaBoost, to segment spine MR images, which achieved good robustness and accuracy, outperforming previous algorithms. Ref. [14] introduced the MRF (Markov Random Field) algorithm to perform well in the segmentation of 3D spine CT images. Subsequently, an integrated algorithm that can integrate multiple medical orientation pictures was developed. Ref. [15] broke the routine and used a single model to segment different orientations of medical images for the first time. Ref. [16] combined FCN and CNN to attain a fully automatic segmentation method with accurate segmentation and considerably enhanced segmentation efficiency. Ref. [17] designed Dense-U-Net, an upgraded network of U-net, which can accurately segment spines without any preprocessing of the original data.

2.2. Research on Medical Image Positioning

In addition to the segmentation of medical images, the location of the lesion is also very important for identifying and diagnosing the disease. It is time-consuming and laborious for doctors to manually mark the location of the lesion, and accurate positioning cannot be guaranteed. An automated positioning method that replaces manual positioning is particularly important. However, due to the particularity of medical images, positioning is relatively difficult. For example, some medical images are more complex and it is difficult to extract features. In addition, there is a possibility that there are deformed lesions, which can easily cause inaccurate positioning.

The research on medical image positioning has also undergone a transformation from semi-automatic to fully automatic. Ref. [18] calculated the probability of the position of the center point of the disc based on MR medical images using a classification tree algorithm

to locate and mark the disc. Ref. [19] combined ASM (shape model) and GVF-snake (active contour model of gradient vector flow) and identified a method for extracting shape features of the disc for the classification of disc degenerative diseases. These methods rely on image morphology. When there are large changes in shape, such as fracture deformities, the positioning effect of the target will be considerably reduced. After 2014, most of the research on the positioning of spine medical images relied on deep neural networks. Ref. [20] proposed an improved J-CNN model based on the CNN model, which can adjust and refine the key point positions generated by the random forest. The use of deep neural networks not only encouraged the transformation of medical image positioning to automation but also greatly improved the speed of positioning. Ref. [21] proposed the use of an improved model based on a deep neural network and verified that deep learning can accelerate the speed of positioning detection under the premise of accuracy. This provides an immense auxiliary effect for clinical diagnosis.

2.3. Research on Classification of Medical Imaging Diseases

Disease classification of medical images is the ultimate goal of medical image segmentation and positioning, which is also the key to intelligent medical-assisted treatment. Regarding the classification of spine-related diseases, as early as 1988, [22] used a multilayer perceptron network to diagnose low back pain and compared it with three groups of doctors at the same time. The results showed that the average diagnosis accuracy rate of the network model was higher than that of the doctors. This is also the first practical application where neural network algorithms achieve effective performance. Ref. [23] proposed a Bayesian classifier-based system to help doctors screen and initially diagnose spinal stenosis.

For the diagnosis of disc herniation, Ref. [24] used the intensity information and texture features of the disc in the MRI to affect the image as a classification condition and constructed five classifiers for the diagnosis of lumbar disc herniation. They are SVM, PCA + LDA, PCA + Naive Bayes, PCA + QDA, and PCA + SVM, and the experimental results show that SVM works best. Compared with Ghosh, [25] added the shape of the disc to the classification condition and proposed a classification framework to achieve degeneration or non-degeneration of the intervertebral disc based on SVM. Ref. [26] proposed a new method for automatically diagnosing degenerative disc disease in the median sagittal plane MRI. This method is based on a machine learning framework and combines the characteristics of different intervertebral discs to locate, segment, and classify lumbar MRI images. Ref. [27] used the AdaBoost model to classify intervertebral disc degeneration on five specific features. Ref. [28] used convolutional neural networks to classify intervertebral discs into six categories, and the model achieved an accuracy of 95.6% in the detection and labeling of intervertebral discs. Ref. [29] proposed a feature fusion method to train an SVM model on the fused feature vector and classify the intervertebral disc.

The above research shows that the classification of spine-related diseases is mainly based on the classification of intervertebral discs. The mainstream classification methods are mainly machine learning SVM classifiers. In recent years, deep learning fusion training methods have gradually appeared. Ref. [7] investigated the effectiveness of three dimensionality reduction techniques and three feature selection techniques on the performance of classifying lumbar spine MRI. Ref. [8] proposed a classifier based on a deep convolutional neural network and considered extensive hyperparameter optimization to improve the performance in intervertebral disc classification. Ref. [9] considered segmentation of the region of interest (ROI) in the disc area to improve classification performance using a convolutional neural network. However, very few studies have focused on the classification of both discs and vertebrae and the sagittal and axial view plane images are used independently, which limits the performance of the deep learning approaches due to the insufficient use of data.

3. Description of Methods

3.1. Mask RCNN Model

Ref. [30] proposed the Mask RCNN model which is based on the Faster-RCNN model proposed by [31]. This model is more effective in detecting small targets. The training process is as follows. First, a fixed-size image is passed to the backbone network which includes ResNet 18, 34, 50, and 101 for feature extraction, and the feature maps are generated using an FPN (feature pyramid network) [32]. Second, the output feature maps pass through an RPN layer for extraction of region proposals. Then, two FCNs (fully connected layer) are used to classify whether the object is the foreground or background and adjust each anchor using bounding box regression. Finally, through the ROI Align method, the region proposal is pooled to the corresponding size and placed into the fully connected layer and the softmax function for classification and bounding box regression. At the same time, the mask branch will segment the target from the image at the pixel level. The framework of Mask RCNN is shown in Figure 1. The specific process is shown below:

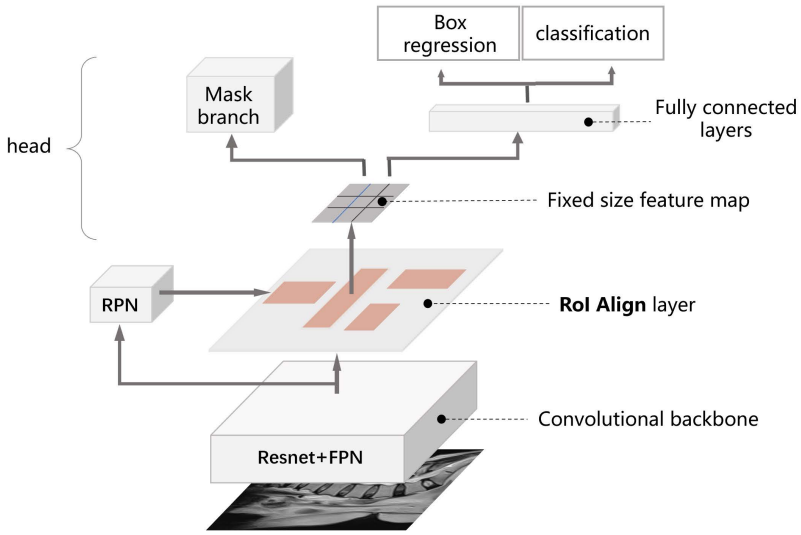

Figure 1. The framework of the Mask RCNN. Mask RCNN uses ResNet and FPN as the backbone network for feature extraction and uses the RPN network to extract target recommendation frames.

1. Using ResNet as the backbone, it will output five feature maps with reduced sizes and different depth features, denoted as C_n, $n = 1, 2, 3, 4, 5, 6$, (where C_1 has a larger size and occupies too much memory, so it will not be used). It is obvious that the deeper the layers, the fewer detailed features of the spatial information indicated by the feature map and the stronger the semantic information displayed.

2. Mask RCNN utilizes FPNs (feature pyramid networks) to make full use of different levels of features, which consist of a feature map M_n attained by a 1×1 convolution operation to modify the channel of C_n and an addition with M_{n+1}. According to this method, M_5, M_4, M_3, and M_2 can be obtained, respectively, as follows:

$$M_n = \begin{cases} \text{Conv}(C_n), n = 5 \\ \text{Upsample}(M_{n+1}) + \text{Conv}(C_n), n = 2, 3, 4 \end{cases} \quad (1)$$

After that, 3×3 convolution blocks are used to eliminate the aliasing effect from upsampling, outputting P_n. P_6 is used for the 5th anchor scale in RPN Generated by subsampling from P_5 with a stride of 2.

3. $P_n, n = 2, 3, 4, 5, 6$ are passed into the RPN to obtain the ROIs. The RPN traverses the feature map with a fixed-sized sliding window whose center position is the anchor. Each anchor has a corresponding scale coordinate in the original image and each of these coordinates can be mapped to the original image according to the scale. Therefore, RPN will output all anchor boxes and then generate positive and negative sample labels through

the classification network, and at the same time, adjust the bounding boxes through the regression method. In this process, the RPN multi-task loss consists of classification loss and bounding box regression loss:

$$L(\{P_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{leg}(t_i, t_i^*) \tag{2}$$

$$L_{cls} = -\log(p_i) \tag{3}$$

$$L_{reg}(t_i, t_i^*) = \sum_i \text{smooth}_{L_1}(t_i - t_i^*) \tag{4}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{5}$$

P_i indicates the probability that the ith anchor is predicted to be the true label, P_i^* is 1 for positive samples and 0 for negative samples, $t_i = \left[t_x, t_y, t_w, t_h\right]$, which represents the bounding box regression parameters of the ith anchor, $t_i^* = \left[t_i^*, t_y^*, t_w^*, t_h^*\right]$ indicates the bounding box regression parameter of the ith anchor corresponding to the GT Box, N_{cls} is the number of all samples in the first mini-batch, and N_{reg} is the number of anchor positions.

4. The RPN network can extract a lot of anchor boxes but not all of them need to be predicted. By screening the top 6000 probabilities and then removing the anchor boxes that exceed the image boundary, 2000 ROIs can finally be selected.

5. Select the feature map P_k from $P_n, n = 2, 3, 4, 5$ and intercept the ROI according to the following formula:

$$k = \left[k_0 + \log_2(\sqrt{wh}/224)\right] \tag{6}$$

where k_0 is the number of layers corresponding to an ROI with an area of 224×224, and w and h are the width and height of each ROI. The formula is easy to understand. Large-scale ROIs should be cut from low-resolution feature maps, which are conducive to recognizing large targets, and small-scale ROIs should be cut from high-resolution feature maps, which are conducive to detecting small targets.

6. Furthermore, classify ROI adjust its position, and perform mask segmentation. The multi-task loss is defined as the same as [30], which is $L = L_{cls} + L_{box} + L_{mask}$. L_{cls} is classification loss, L_{box} is bounding-box loss, and L_{mask} is mask loss.

3.2. VGG

In the 2014 ImageNet Challenge, Ref. [33] achieved first and second places in the localization and classification tracks, respectively, which presented a significant aspect of CNNs: their depth. The VGG model takes 11–19 weight layers to ensure learning of more complex patterns. To push the depth to 11–19 and also at a smaller cost, the VGG model uses two convolutional kernel sizes of the same size (3×3) instead of the previous convolutional kernel size (5×5). In other words, the VGG model accomplishes the learning task of the same size perceptual field by overlaying two hidden layers with a size of 3×3, and this module enhances the network depth and improves the learning ability of the network to a certain extent.

3.3. ResNet

The deep convolutional residual learning network proposed by [34] won first place in the ImageNet competition, and the error rate of the network in the competition reached 3.57%. The proposal of the ResNet network solves the problem in which the number of network training layers cannot be deeper. Compared with ordinary networks, ResNet has many more

shortcut paths. Its network is a block structure, which is very easy to modify and expand. By adjusting the number of channels in the block and the number of stacked blocks, you can easily adjust the width and depth of the network. In this way, networks with different expression capabilities can be obtained without worrying too much about the degradation of the network. As long as the training data are sufficient and the network is gradually deepened, better performance can be obtained.

3.4. MobileNet

Ref. [35] introduced MobileNetV1 in 2017 as a lightweight CNN network with a focus on mobile and embedded devices. The depthwise convolution proposed by this network significantly reduces the amount of training and the number of parameters of the model, compared to, for example, VGG networks and ordinary convolutional networks, although there is a small decrease in accuracy, but the network efficiency is greatly enhanced. Compared to MobileNetV1, MobileNetV2, which was proposed by [36], has a smaller model with more accuracy, and the inverted residual and linear bottleneck modules have been introduced to the network. Ref. [37] then adjusted the bottleneck of inverted residuals in MobileNetV2, applied a neural architecture search approach to search for parameters, re-analyzed the running duration of each layer in the network, and reconfigured the time-consuming layers. In terms of precision and latency, MobileNetV3 Large and MobileNetV3 Small are superior to MobileNetV2.

4. Experiment and Design

In brief, the logic flow of the proposed two-stage classification process is shown in Figure 2. In this study, we initially employed the Mask RCNN model to localize and classify the lumbar spine's intervertebral discs and vertebrae in sagittal spine images. Although we used masks parallel to the image to cover the regions of interest, misalignment between the masks and the actual regions of interest could potentially affect the classification results. However, by using the center points of the bounding boxes as the locational references, the deviations were kept within acceptable limits.

Subsequently, we employed quadratic spline interpolation to further refine the spinal curve based on the preliminary localization results. This process not only facilitated more precise center point determination but also compensated for any predictive points that the Mask RCNN may have missed. The accurately fitted spinal curve allowed for sagittal slices along the spinal axis to be more perpendicular to the spinal curve, effectively covering the target area and reducing irrelevant noise interference.

In processing the intervertebral discs, we proportionally trimmed the sagittal slices of the discs and their corresponding central axial slices, and then vertically concatenated them to create multi-angle views of the discs. This not only increased the features available but also significantly enhanced the accuracy of disc classification. For the vertebrae, we directly used the trimmed sagittal slices for model training.

Finally, the multi-view images of the discs and the sagittal slices of the vertebrae were fed into various classification networks including ResNet, VGG, and MobileNet for feature extraction and further classification. These steps ensured the ability to evaluate and classify from multiple perspectives, thereby enhancing the overall robustness and accuracy of the model.

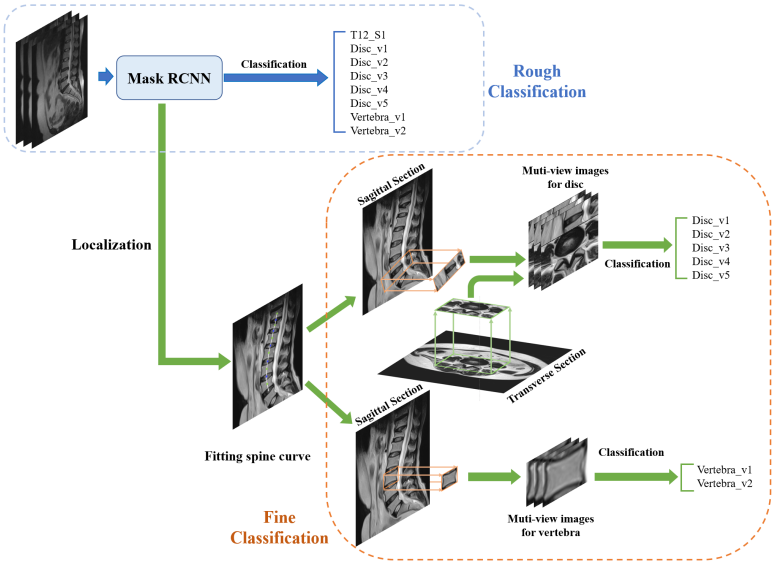

Figure 2. Research method for the rough and fine classification process.

4.1. Spine Dataset

The spinal data for this experiment come from the "Alibaba Cloud Tianchi Algorithm Competition "Spark" Digital Human AI Challenge-Intelligent Diagnosis of Spine Diseases", which is provided by Wanli Cloud and AllinMD Orthopaedics. It provides T1 and T2 sagittal images of the spine and T2 axial position (FSE/ TSE) images, and the picture format is DCM. Figure 3 displays example plots of T1-weighted sagittal, T2-weighted sagittal, and T1-weighted axial images in the dataset. We chose T2-weighted imaging for training our model primarily for its high efficacy in depicting edema and inflammation around intervertebral discs, which enhances the identification and classification of pathological features. During the detailed classification phase, composite images that merged both axial and sagittal views were exclusively created using T2-weighted imaging. This ensured uniformity in imaging technique across the merged views, optimizing our model's learning process.

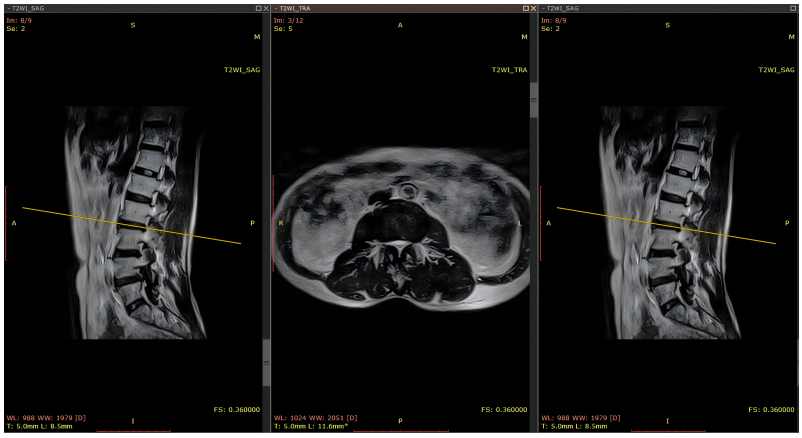

Figure 3. The example types of images in the dataset provided by Tianchi competition. (**Left**) T2-weighted sagittal image. (**Middle**) T2-weighted axial image. (**Right**) T1-weighted sagittal image. T1-weighted images are produced by using short TE and TR times, while T2-weighted images are produced by using longer TE and TR times.

The data contain 150 pieces of labeled training set data, 50 pieces of validation set data, and 50 pieces of unlabeled test set. The key points are marked in the sagittal position of the T2 intermediate frame in the data. Each marked lumbar vertebra contains six intervertebral

discs and five vertebrae (without missing values), of which six intervertebral discs are: T12-L1, L1-L2, L2-L3, L3-L4, L4-L5, L5-S1. The six vertebrae are L1, L2, L3, L4, L5. The annotations provided by the competition are in JSON format, which contains key point information, lesion types, and Dicom image information. The purpose of this experiment is to classify the discs and vertebrae of the lumbar spine. The vertebral bodies are divided into two types, and the intervertebral discs are divided into five types. Table 1 shows the disease categories of intervertebral disc and vertebra.

Table 1. Disease category of intervertebral disc and vertebra.

vertebra	
coding	**type**
V1	Normal
V2	Degeneration
disc	
coding	**category**
V1	Normal
V2	Bulge
V3	Protruded
V4	Extruded
V5	Schmor

Dicom medical imaging is one of the most popular medical influences, bringing great convenience to the exchange of information in the medical field. M. Mustra (2008) [38] evaluated that Dicom's influence can meet the needs of CT, MRI, SPECT, PET, and other different imaging equipment imaging data transfer and exchange. The exchange is conducive to image storage and has high resolution, which can save a lot of medical information and avoid the confusion and redundancy of binders. We can use Matlab, Python, C++, and other software to read tags in Dicom images. Tags include four categories: Patient Tag, Study Tag, Series Tag, and Image Tag. These four tags can help researchers obtain most of the information about the image. Table 2 shows several tags and their functions involved in this experiment.

Table 2. Tags and the functions commonly used in this article.

Tag	**Function**
(0020, 000d)	Study Instance UID
(0020, 000e)	Series Instance UID
(0008, 0018)	SOP Instance UID
(020, 0032)	Image Position (Patient)
(0020, 0037)	Image Orientation (Patient)
(0028, 0030)	Pixel Spacing

Study Instance UID, Series Instance UID, and SOP Instance UID are the only indicators that can determine the examination and sequence.

Image Position is the world coordinate of the upper left corner of the spine image, with a total of three parameters. The upper-left corner of the image is the origin of the image coordinate system.

Image Orientation, which contains three parameters, indicates the direction cosine of the initial row and column relative to the patient. The orientation of the patient determines the direction of the axes, with the x-axis increasing toward the left side of the patient, the y-axis increasing toward the back of the patient, and the z-axis increasing toward the head of the patient. Therefore, it is able to define the link between the image coordinate system and the anatomical coordinate system and identify the position of the image in the anatomical coordinate system. Combining Image Position and Image Orientation can

calculate the position of the picture relative to the human body and the position of the axial bitmap, sagittal image, and coronal image in the world coordinate system.

Pixel Spacing represents the interval between each pixel in the world coordinate system, which corresponds to a distance of 0 in the image coordinate system.

According to the three tag indicators of Image Orientation, Image Position, and Pixel Spacing, any conversion between the world coordinate system, anatomical coordinate system, and image coordinate system can be completed.

4.2. Dataset Division

We divide the data into a training set, a validation set, and a test set. Since the public test set of the competition does not include labeling and classification, it cannot be used to judge the model, so we randomly divide the verification set provided by the competition into two parts: the verification set and test set of this experiment. The distribution of the three sets is shown in Table 3.

Table 3. Distribution of training set, validation set, and test set.

Type	Quantity
Train	150
Val	26
Test	25

A piece of data includes 35 to 70 medical images in Dicom format, each of which contains different sagittal, coronal, and axial three-dimensional images of different frames, only one middle frame of sagittal image is annotated. In the labeled information, there are a small number of vertebral bodies or intervertebral discs containing two categories. In this experiment, only the previous category is selected for classification, and multiple classifications of the same target are not involved.

4.3. Key Point Location and Rough Classification Based on Mask RCNN Model

4.3.1. Mask Design

Mask draws on the production of semiconductors in the industry, using photolithography in the production process. The location only corrodes or diffuses the area outside the designated area, thus introducing the concept of the "mask" to cover the selected area with an opaque template. In this experiment, the role of the mask is to extract the region of interest. First, set the mask to a matrix of all 0s with the same size as the picture, set the range of the region of interest, set the matrix value in the range to 1, and set the rest to all 0. Each target corresponds to a mask matrix, for example, for a sagittal image containing 11 vertebral bodies and disc marks, a total of 11 mask images are generated.

There are many ways to make a mask. One of the common ways is to use image annotation tools such as VGG Image Annotator (VIA) and Labelme for manual annotation. It can use rectangular, circular, polygonal, or dotted borders for annotation, and it can simultaneously generate classification labels that can be directly used for training. Using graphical annotation tools can better mark the area of interest, but this method is time-consuming and labor-intensive. This article directly uses the marked point information provided by the data to generate a rectangular box as a mask, which is more convenient.

This experiment consists of nine classification tasks. The nine categories are (1) BG (background); (2) disc_v1; (3) disc_v2; (4) disc_v3; (5) disc_v4; (6) disc_v5; (7) vertebra_v1; (8) vertebra_v2; (9) T12_S1. Among them, T12_S1 is specifically used to define the boundaries of the entire lumbar spine, preventing the Mask RCNN algorithm from incorrectly identifying vertebrae and discs outside the designated lumbar area.

This experiment is centered on the marked points of the intervertebral disc and vertebra and constructs a horizontal rectangular frame with width and height of $36/512 \times$ width and $26/512 \times$ height (width and height are the width and height of the sagittal figure) as

the mask of discs and vertebra. A sagittal position is randomly selected from the training set, and the mask is displayed on the sagittal position map, as shown in Figure 4.

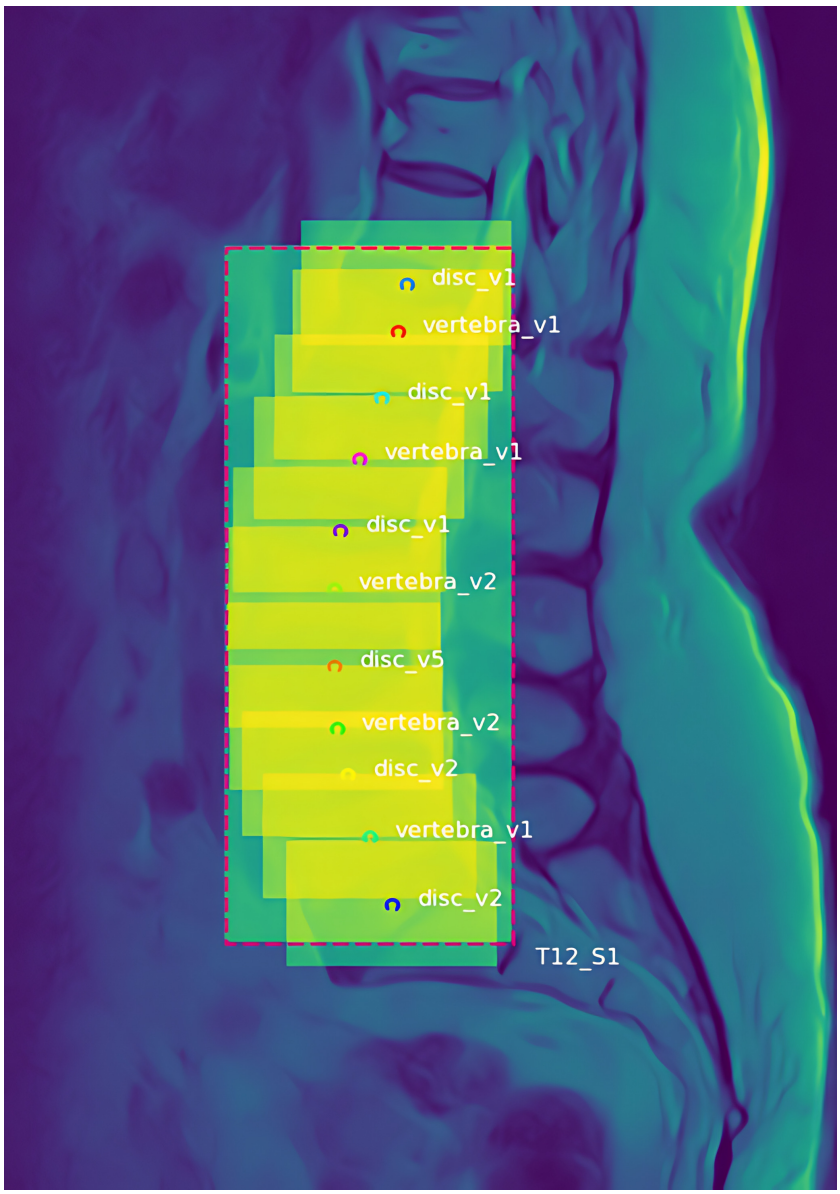

Figure 4. Masks in sagittal images. Each sagittal image has 12 masks: one mask for the full lumbar spine (T12_S1) localization and 11 masks for the localization of all discs and vertebrae.

4.3.2. Spine Detection

For the positioning of the vertebra and intervertebral disc, the adjacent vertebral bodies and intervertebral discs can be positioned by using the position of the upper and lower-vertebrae or intervertebral discs, but this method is not feasible in actual engineering practice. In fact, there are missing annotation points on the spine image, which will cause if a vertebra or disc fails to be located, all subsequent predictions will be wrong. In this paper, the method of image location is used to transfer the middle frame of the sagittal image of the MRI image into the fine-tuned Mask RCNN model to locate and classify all the intervertebral discs and vertebrae as a whole. Target detection of the spine is divided into positioning tasks and classification tasks. The input is a spine picture, and the output is a prediction of the position of the disc and the vertebra in the spine picture and its corresponding category. The positioning task is to input the MRI image of the spine and output the coordinate position information of the vertebra or the disc. The

classification task is to input an MRI image of the spine and output a feature vector. This feature vector represents the probability that the image belongs to each category. The size of the probability can determine which category the image belongs to. Therefore, the target detection task is to output not only a coordinate position information but also a feature vector of classification information.

Take the recognition of a sagittal image as an example. When a spine image is input a into a multi-layer convolutional neural network, the network will output a feature map, them classify the feature map through the fully connected layer, and finally obtain the feature vector. This feature vector will be composed of two parts, one is the coordinates and the other is the one-hot code representing the category and a credibility value p, which is 0-1. The coordinates are composed of center coordinates (x, y) and width and height (w, h). One-hot codes are used to indicate categories. The discs in the lumbar spine are divided into five categories, and the vertebral bodies are divided into two categories. In this paper, five types of intervertebral discs and two types of vertebral bodies are classified together, plus background and T12_S1 categories, a total of nine9 categories. The codes are 0 0000 0001, 0 0000 0010, 0 0000 0100, 0 0000 1000, 0 0001 0000, 0 0010 0000, 0 0100 0000, 0 1000 0000, 1 0000 0000, respectively. The combined label is $\mathrm{Z} = (\mathrm{x}, \mathrm{y}, \mathrm{w}, \mathrm{h}, 0, 0, 0, 0, 0, 0, 0, 1, \mathrm{p}), \ldots,$ $(\mathrm{x}, \mathrm{y}, \mathrm{w}, \mathrm{h}, 0, 0, 0, 0, 0, 0, 1, 0, \mathrm{p})$, $(\mathrm{x}, \mathrm{y}, \mathrm{w}, \mathrm{h}, 1, 0, 0, 0, 0, 0, 0, 0, \mathrm{p})$. With this label template, the real label and the output of the model can be used to calculate the loss during training and backpropagate after the loss is output, and the coordinates, category, and credibility predicted by the model can be obtained during testing. With the predicted coordinates, the target can be marked on the image.

4.3.3. Model Training Configuration

In this experiment, suitable adjustments are made to the configuration of the Mask RCNN training. Table 4 shows most of the model training hyperparameters in this experiment. All of the images have been scaled down to 512 pixels by 512 pixels. The number of classification classes, which is denoted by the parameter NUM_CLASSES and is set to 9, consists of the following: one class for the background, one class for the full lumbar spine (T12_S1), two classes for the vertebra categories, and five classes for the disc categories. We use ResNet 18, 34, 50, and 101 as the backbone networks, and each network undergoes training for a total of 100 epochs.

Due to the unique challenges associated with identifying vertebral bodies and intervertebral discs, we did not use a pre-trained model in this study. Compared to general image recognition tasks, these anatomical structures display distinct and limited features. The use of non-specific datasets may lead to overfitting to non-relevant features, which may obscure critical medical details and specific pathological characteristics that are essential for accurate medical imaging analysis. Further, reliance on general datasets for pretraining might impair the model's ability to generalize, leading to suboptimal performance on real medical images.

We simply choose the learning rate as 0.001. ALIDATION_STEPS is the number of validation steps that should be run at the end of each training epoch, and STEPS PER EPOCH is the number of training steps that should be executed throughout each epoch. In spite of the fact that setting them to large numbers can result in more accuracy, we decided to set them to 100 and 30, respectively, because the dataset and epoch were rather small. This allowed us to strike a balance between accuracy and training efficiency. We give the TRAIN_ROIS_PER_IMAGE hyperparameter the value of 200, which indicates that 200 ROIs per image are fed to the classifier or mask head. The DETECTION_MIN_CONFIDENCE parameter is set to 0.8, which represents the minimal probability value required to detect instances, so regions of interest (ROIs) that fall below this threshold are skipped. DETECTION_NMS_THRESHOLD is a non-maximum suppression threshold for detection, which is set to 0.3 to discard results that exceed this threshold, to ensure that only structures within the lumbar region are analyzed. Except for the above hyperparameters, the other hy-

perparameters in the experiments are kept consistent with those in the original experiments of Mask RCNN.

Table 4. Model training hyperparameters.

Hyperparameters	Value
Image Size	512 × 512
Classification classes	9
Epoch	100
Learning rate	0.001
ALIDATION_Steps	100
Steps	30
ROIs	200
Detection Min Confidence	0.8
Detection NMS Threshold	0.3
RPN_NMS_Threshold	0.7
Mini_Mask_Shape	(56, 56)
RPN_Anchor_RATIOS	[0.5, 1, 2]

4.3.4. Experimental Platform

This experiment uses Keras and TensorFlow frameworks. The hardware platform of this experiment is AMD3600 CPU, the GPU is a GTX1660super, and 32 GB RAM and a 500 G hard disk were used. The software platform is Windows 10, and the compilation platform is PyCharm 2021.1.3.

4.4. Classification of Multi-Angle View of Disc and Vertebra Classification Based on ResNet Network

This part aims to improve the phenomenon of poor classification effect in the positioning and rough classification using the Mask RCNN model. Different from making a parallel mask, in this part we fit the spine curve to make a rectangular box parallel to the target and cut the slice. The sagittal slices and axial slices of the intervertebral disc are spliced up and down to form a multi-view image of the intervertebral disc, which is put into four kinds of ResNet networks for training. In this part, the vertebrae and discs are trained separately to prevent them from affecting each other.

4.4.1. Spine Curve Fitting

When using Mask RCNN for key point prediction and classification, a horizontal rectangular mask with key points as the center is made using the information of the marked points. Although this covers most of the characteristic information of vertebrae and intervertebral discs. However, the spine has a certain physiological curve, and the direct use of the horizontal mask will lose its features to a certain extent. The vertebrae and discs are not parallel to the rectangular frame, and the calculator will extract the wrong features. Using quadratic spline interpolation to fit the spine curve can solve this problem.

Common curve fitting methods include Lagrangian interpolation, Newton interpolation, piecewise interpolation, polynomial curve fitting, and so on. It has been verified that due to the few key points of the labeled data and the small curvature of the spine curve, the use of more complex high-order polynomial interpolation functions or complex functions is likely to cause the Runge phenomenon, whereas the use of simple interpolation functions such as linear interpolation to fit the spine curve. It is not smooth enough, so it is more reasonable to choose the quadratic sample question interpolation method to fit the spine curve. However, the spine curve fitted by simple interpolation functions such as linear interpolation is not smooth enough, so it is more reasonable to choose the quadratic spline interpolation method to fit the spine curve.

The quadratic spline difference is a piecewise interpolation method. The interval [a, b] is divided into n subdivisions, and each segment of the divided interval is fitted with a quadratic polynomial. In this experiment, there are 12 marking points for the intervertebral discs and

vertebral bodies (take the sagittal markings without missing values as an example). Since the middle area of the lumbar spine is basically vertical, the abscissas of different vertebral bodies or intervertebral discs are the same, so the quadratic function cannot be solved. We can exchange the abscissas and ordinates of the marked points to solve this problem. The specific steps of the quadratic spline difference method to fit the spine curve are as follows:

1. Mark the coordinates of the marked point as following:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots (x_{i+1}, y_{i+1}) \quad i = 11 \tag{7}$$

2. Using the ordinate as the basis for dividing the interval, divide the interval [a, b] into 12 sub-intervals:

$$\begin{aligned} [a, y_2], [y_2, y_3], [y_3, y_4] \ldots [y_i, b] \quad i = 11 \\ a = y_1, b = y_{i+1} \end{aligned} \tag{8}$$

3. Construct a quadratic interpolation function in each interval, denoted as $S_i(y)$:

$$x_i = a_i + b_i y_i + c_i y_i^2 \quad i = 1, 2, \ldots 11 \tag{9}$$

4. All known points meet the following conditions:

$$\begin{aligned} S(y_i) &= x_i \\ S_i(y_{i+1}) &= x_{i+1} \\ S_{i+1}(y_{i+1}) &= x_{i+1} \end{aligned} \tag{10}$$

The quadratic interpolation function of two adjacent intervals has a continuous first-order derivative and continuous second-order reciprocal at the intersection point:

$$\begin{aligned} S_i'(y_{i+1}) &= S_{i+1}'(y_{i+1}) \\ S_i''(y_{i+1}) &= S_{i+1}''(y_{i+1}) \end{aligned} \tag{11}$$

According to the above conditions, simultaneous equations can calculate a_i, b_i, c_i the parameters of 12 interpolation functions.

Figure 5 is the curve-fitting diagram of the spine. The rectangular frame is intercepted according to the normal and tangent of the spine fitting curve at the key points to ensure that the mask can completely cover the disc and the vertebra and that the direction is level with the mask.

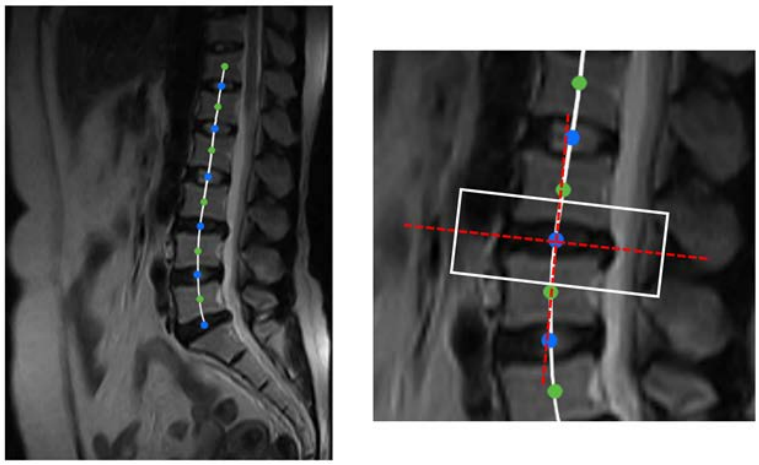

Figure 5. Spine curve fitting and section cutting. (**Left**) A smooth spine curve was obtained by fitting the spine curve using the quadratic spline difference method. (**Right**) A rectangular slice of size (50/521*W, 150/521*H) is cut along the tangent direction of the fitted curve.

4.4.2. Axial Bitmap Positioning of Disc

In order to simulate the doctor's diagnosis process and ensure the credibility of the diagnosis, the axial image feature is added. This can increase the accuracy of the prediction. By transforming the coordinates of the space coordinate system and the image coordinate system, a CT/MR medical image positioning line can be drawn. Medical imaging includes three coordinate systems: world coordinate system, anatomical coordinate system, and image coordinate system. Through the conversion between these three coordinate systems, the position of the pixel in the image coordinate system in the world coordinate system and the position of the sagittal, coronal, and axial positions in the world coordinate system relative to the human body and the instrument can be obtained.

In this experiment, the method of drawing the positioning line of the axial position of the intervertebral disc is as follows:

1. First, filter out all the axial bitmap images. Read the six parameters in Image Orientation, cross-product the vector composed of the first three parameters and the vector composed of the last three parameters, and the vector obtained is the normal vector of the image. If the normal vector points to the x-axis (x $\neq$ 0, y = 0, z = 0), it means that the image is a sagittal image; if the normal vector points to the y-axis (x = 0, y $\neq$ 0, z = 0), it means that the image is a coronal image. If the normal vector points to the z-axis (x = 0, y = 0, z $\neq$ 0), it means that the image is an axial image.

2. Read the tag = ["0020 | 0032", "0020 | 0037", "0028 | 0030"] field to obtain the Image Position, Image Orientation, Pixel Spacing, and Image Position of the positioning map (sagittal image) to obtain the coordinate value of the upper-left corner of the image in the space coordinate system, denoted as P0. Image Orientation has six parameters, denoted as $O_1 = (a, b, c, d, e, f)$, where $r_1 = (a, b, c)$ and $c_1 = (d, e, f)$. Pixel Spacing represents the width and height of each pixel in the space coordinate system in the actual image coordinate system, denoted as $d = (d_1, d_2)$. The width and height of the sagittal bitmap in the image coordinate system are w and h. According to the above information, the coordinates P_1, P_2, P_3, P_4 of the upper-left corner, upper-right corner, lower-right corner, and lower-left corner of the sagittal bitmap in the spatial coordinate system can be obtained, which are:

$$\begin{aligned} P_1 &= P_0 \\ P_2 &= P_1 + \vec{r} \cdot (d_1 \cdot w) \\ P_3 &= P_2 + \vec{c}(d_2 \cdot w) \\ P_4 &= P_1 + \vec{c}(d_2 \cdot h) \end{aligned} \tag{12}$$

3. In the same way, the coordinate values of the upper-left corner, upper-right corner, lower-right corner, and lower-left corner of the axial image in the space coordinate system can be calculated.

4. Calculate the vertical vector from the upper-left corner, upper-right corner, lower-right corner, and lower-left corner of the axial position in the space coordinate system to the sagittal position, denoted as D_1, D_2, D_3, D_4. If the four values have different signs in pairs, it means that the axial bitmap and the sagittal bitmap intersect. Cross multiplying the normal vector of the axial bitmap and the sagittal bitmap to obtain the vector of the intersection line, which can be drawn with the sagittal bitmap. Figure 6 is the line map of the axial position of the intervertebral disc.

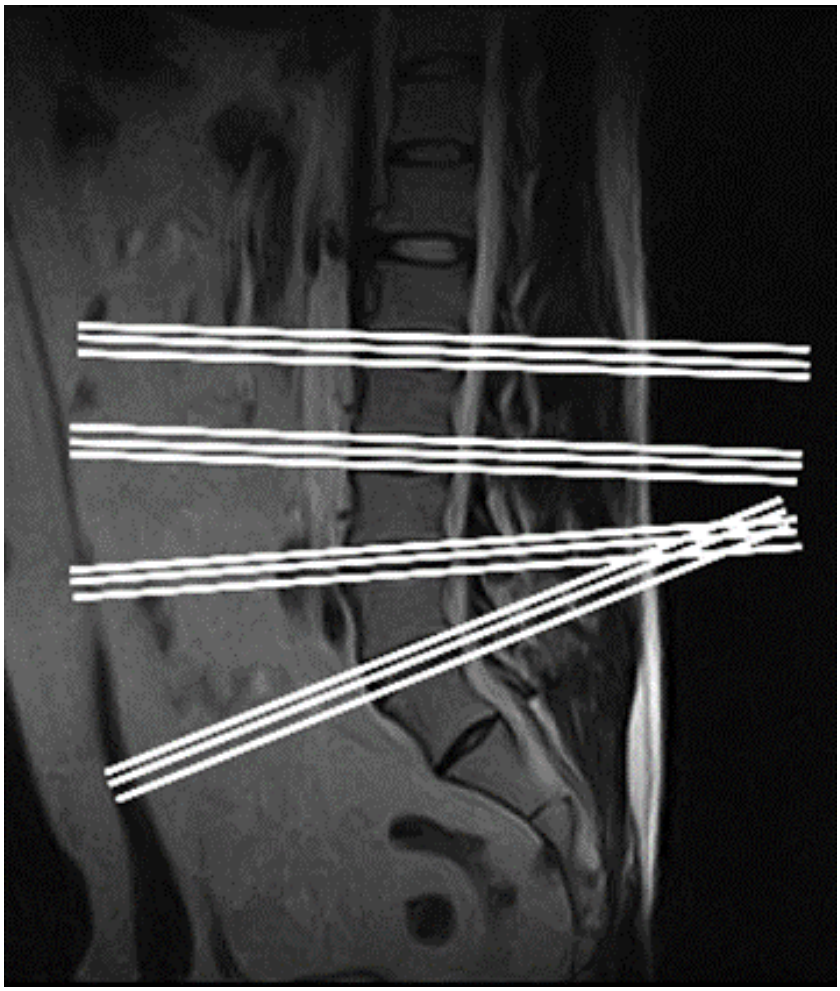

Figure 6. Axial positioning line map of the disc. This image is one of the T2-weighted sagittal images, and the white lines are the projection of the axial images in the sagittal image. In this example, there are three corresponding axial views of each disc.

After locating all the axial bitmaps in the sagittal image, it is necessary to find the axial bitmap that best represents the features at the key point, so the axial bitmap closest to the key point should be selected. Calculate the distance from the key point to all the axial bitmaps and compare them to obtain the axial bitmap with the closest distance to the key point. Since the axial bitmap corresponding to the annotation information in the data is not complete, there is no corresponding axial bit slice at a small part of the mark point position, so the distance between the key point and the axial bitmap should be restricted by a distance < 4 (through observation, the axial slice that can fully represent the features at the key point is about 4 pixels above and below the key point), avoiding selection of an axial bitmap that is closest to the marked point but is not corresponding to the marked point. Thus, the axial bitmap of the corresponding position of each key point can be obtained.

4.4.3. Multi-Angle View of Disc

Ref. [39] once put the intervertebral disc sagittal slice and the intervertebral disc axial bitmap into two identical feature extraction networks and classification networks for disc segmentation and lesion recognition. Considering that the sagittal slices at key points have a certain correlation with their corresponding axial bitmaps, putting them into the feature extraction network and the classification network training will lose the correlation information of the two. In this paper, the sagittal slice of the intervertebral disc is spliced with the axial position of the intervertebral disc, retaining its relevant characteristics.

Ref. [40] focused on the morphological analysis and feature recognition of the herniated intervertebral disc during the classification and geometric diagnosis of intervertebral disc herniation (HIVD). It shows that the diagnosis of degenerative disc disease is inseparable from the discrimination of disc herniation. In order to ensure that the model is not disturbed by unnecessary features, the features are located in the area of the intervertebral disc protrusion. According to [41], disc diseases are mainly characterized by the circumference in the axial section. For example, bulging is defined as asymmetric if it is more evident in one section of the periphery of the disc but is not so focal as to be characterized as a protrusion. In our dataset, it can be found that the intervertebral disc protrusion is primarily located in the middle of the lower portion of the axial section image. Based on this, we suppose the height of the axial image is h and the width is w, and take $(9/14 \times h, 1/2 \times w)$ as the center point and intercept a rectangular box with a height and width ratio of 1:1.5. Suppose the height of the sagittal image is H and the width is W. Take the key point as the center and divide it according to the ratio of height to width of 1:3. The width is $50/512 \times W$ and the height is $150/512 \times H$. The axial image and the sagittal slice at the interception are spliced up and down, and the width and height are unified to 224×224 to obtain a multi-angle view of the disc. Figure 7 shows the splicing process of the multi-angle view of the disc.

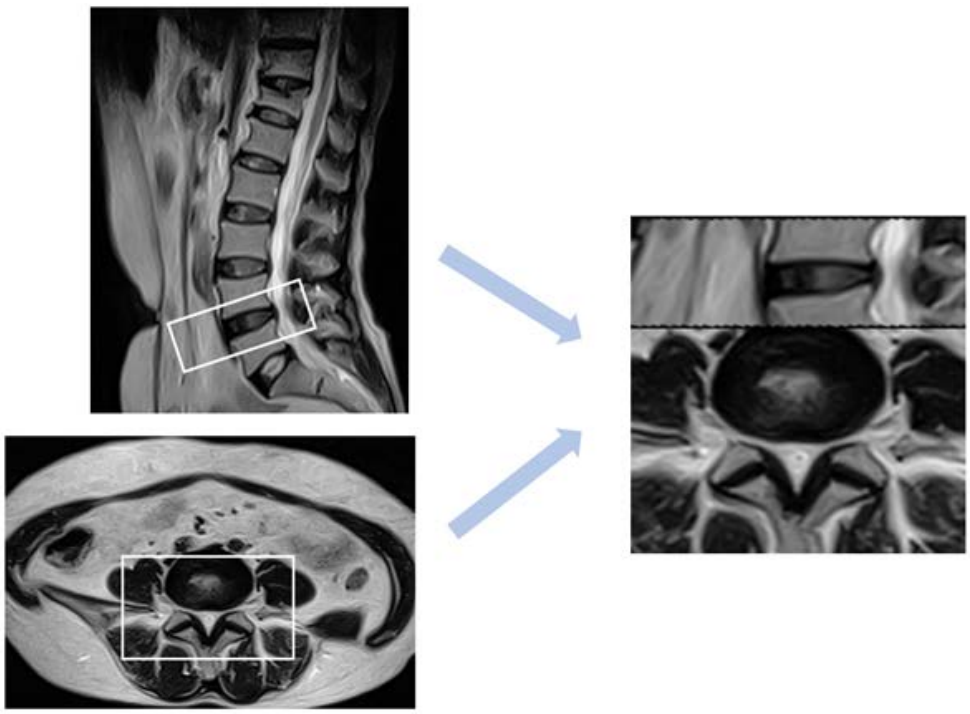

Figure 7. The splicing method of a multi-angle view of the disc. The sagittal and axial slices are stitched together top and bottom to form a 224×224 multi-view image of discs.

4.4.4. Vertebra Slice

The cutting method of the vertebra is similar to that of the intervertebral disc. It is also centered on the key point and cut along the tangent and normal directions of the spine fitting curve. The width and height ratio is set to 1:1, and the width and height are both $55/512 \times W$. Figure 8 is a schematic diagram of the vertebra slice. It can be seen from the figure that the vertebra and the rectangular frame are basically parallel.

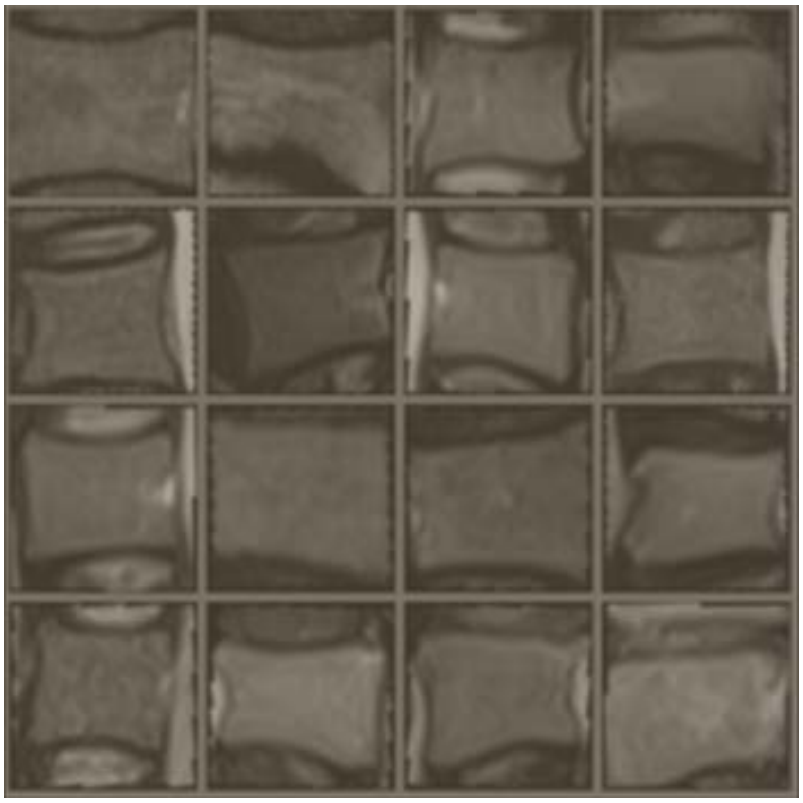

Figure 8. Examples of vertebra slices. The vertebra and the rectangular frame are basically parallel.

5. Results and Discussion

5.1. Performance Metrics

We use precision, recall, and F1 scores as performance metrics for the fine classification of vertebrae and discs. The classification of vertebrae is a dichotomous task, so the value of its F1 score is the harmonic mean of precision and recall, while the classification of discs is a multi-class task, so we use the micro-average method to calculate its F1 score. The related formulas are shown in Table 5.

Table 5. Performance metrics.

Performance Metrics	Formulas/Meanings
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1 score	$2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
Micro_Precision	$\frac{\sum_{i=0}^{n} TP_i}{\sum_{i=0}^{n} TP_i + \sum_{i=0}^{n} FP_i}, n = 4$
Micro_Recall	$\frac{\sum_{i=0}^{n} TP_i}{\sum_{i=0}^{n} TP_i + \sum_{i=0}^{n} FN_i}, n = 4$
Micro_Average F1 score	$2 \cdot \frac{Micro\_Precision \cdot Micro\_Recall}{Micro\_Precision + Micro\_Recall}$

5.2. Results of Key Point Location and Rough Classification Based on Mask RCNN Model

This experiment tested the network performance of the dataset in Mask RCNN. Table 6 is the F1 score and mAP index obtained when the backbone is set to ResNet18, ResNet34, ResNet50, and ResNet101 networks.

The experimental results show that the mAP and F1 score values of ResNet 18 are basically higher than those of other networks with deeper network layers. The overall accuracy rate is between 0.62 and 0.71, the recall rate is around 0.45, the mAP is around 0.53 to 0.58, and the F1 score value is between 0.4–0.45.

Table 6. Mask RCNN model performance.

Model	Precision	Recall	mAP	F1 Score
ResNet 18	0.6839	0.4585	0.5706	0.4455
ResNet 34	0.7146	0.4041	0.5353	0.4027
ResNet 50	0.6981	0.4583	0.5747	0.4421
ResNet 101	0.6237	0.4833	0.5463	0.4371

As shown in the Figure 9, the training loss curve and validation loss curve decline continuously with a similar trend until stabilizing at a value that is virtually identical. There is neither overfitting nor underfitting, as the generalization gap between them is quite small.

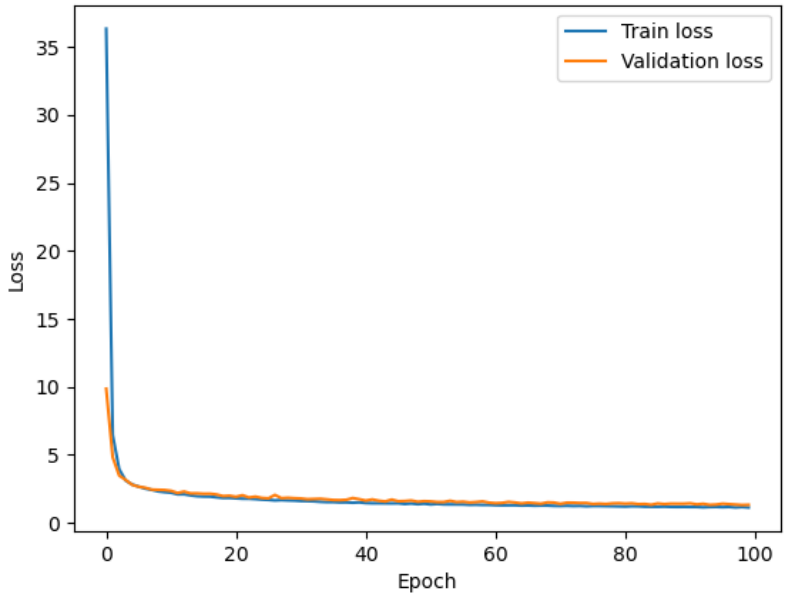

Figure 9. Learning curves for Mask RCNN model with ResNet50 backbone.

Figure 10 presents a prediction output map randomly selected in the test set. The prediction outputs the position coordinates x and y of the prediction point and also outputs the class of each prediction point and the probability p of the class. Through observation, it can be found that the positioning points in the figure are relatively accurate, the prediction points are basically in the middle area of the vertebra and the disc. The positioning task is highly completed. However, because there are still intervertebral discs and vertebral bodies that are not part of the lumbar spine, and their characteristics are similar. Therefore, it can be found that a small number of discs and vertebrae other than T12_S1 are also located and predicted, so the added T12_S1 category can be used to exclude non-lumbar discs and vertebral bodies, making the positioning more accurate.

Figure 11 shows a grid of ground truth objects and their predictions randomly selected from a test set. In the figure, the abscissa is the true category label, and the ordinate is the predicted category label. As you can see in the figure, T12_S1, and the five vertebrae are correctly predicted, but the classification of the intervertebral disc is not accurate enough, and the classification task is generally poorly completed. The reason for the poor classification of intervertebral discs is that the range of the discs is small and flat. When the degree of curvature of the spine is large, if the mask size is set to completely cover the intervertebral discs, many other features and noises besides the discs will be generated, which will interfere with

classification. Secondly, the lesion features of the disc are small and difficult to detect. This is also the reason why the discs and the vertebrae were subdivided in the follow-up experiment.

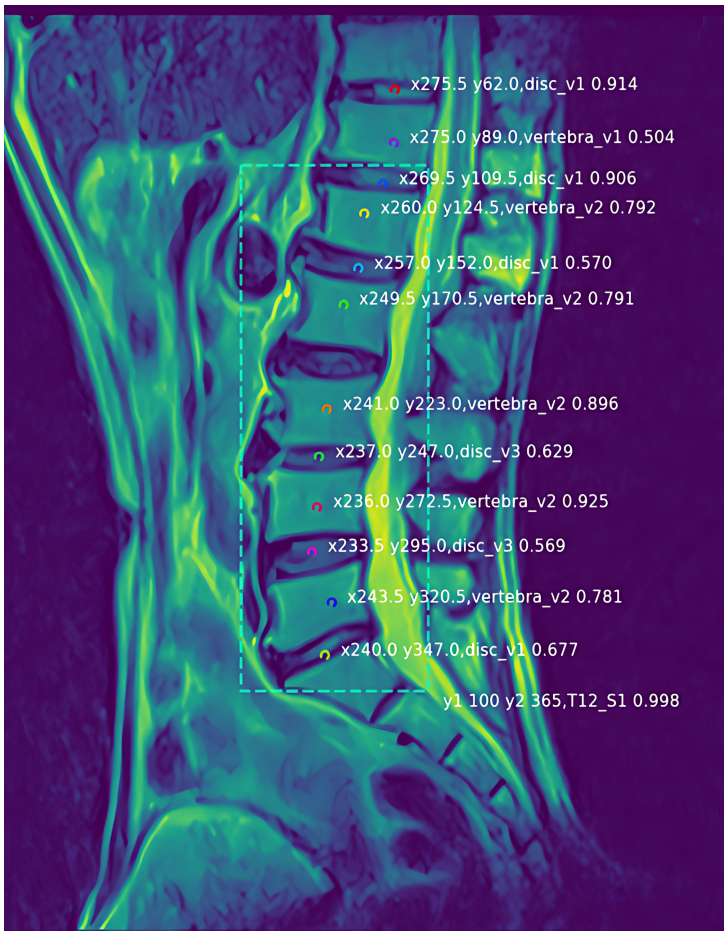

Figure 10. Prediction results of the Mask RCNN model. It outputs the name and category of each disc, vertebra, and entire lumbar spine, together with their horizontal and vertical coordinates in the sagittal images.

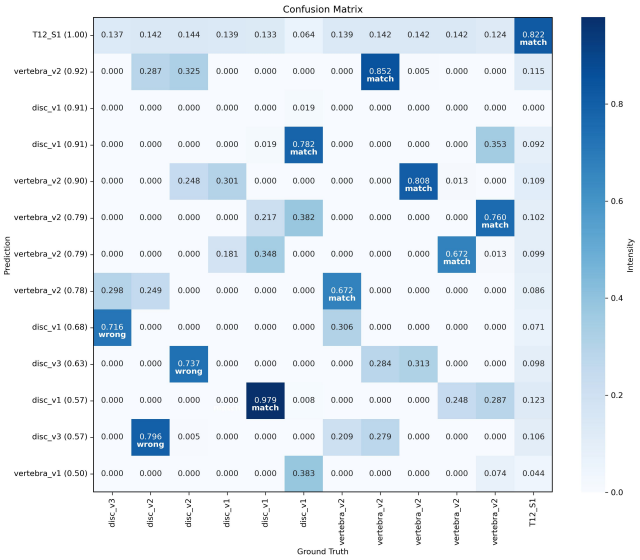

Figure 11. Grid of ground truth objects and their predictions. The model achieved a high accuracy in predicting the localization of T12_S1 and a confidence value of 0.82 for the determination of its type.

5.3. Classification Results of Multi-Angle View of Disc

In this section, through a thorough comparison, we expect to verify that using the synthesized multi-angle view of the intervertebral disc, the classification accuracy of different deep-learning methods will be greatly improved compared with the cases of using sagittal and axial slice images independently. We used ResNet, MobileNet, VGG, and their derivative networks for classification. To avoid model overfitting due to the small dataset, two regularization techniques, batch normalization and dropout, were added in the last two layers of each network. We used precision, recall, and F1 scores as the performance metrics of the model. In addition, we compared the classification results of the intervertebral disc multi-view images with those of the sagittal image only and the axial image only. The results of the comparison are displayed in Table 7, where it can be seen that using multi-view images for classification produced roughly superior results on every same network. Specifically, the multi-view image classification of intervertebral discs showed the best results in the VGG16 network with an F1 score of 96.36%, much higher than the highest F1 score of 86.48% obtained in MobileNetV2 for the sagittal image only and 83.15% for the axial image only in the VGG16 network. Moreover, limited by the size of the dataset and the resolution of the images, better results are shown when using the network with fewer layers, regardless of which kind of disc images are used.

Table 7. Comparison of disc classification results using different types of images: multi-view images, sagittal image only, and axial image only.

	Multi-View Image			**Axial Image Only**			**Sagittal Image Only**		
Model	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
ResNet18	0.9000	0.9333	0.9018	0.7979	0.7991	0.7972	0.7995	0.7248	0.7623
ResNet34	0.8583	0.8667	0.8138	0.8390	0.7769	0.8011	0.7591	0.7803	0.7618
ResNet50	0.8800	0.9000	0.8646	0.7964	0.8060	0.7999	0.7988	0.8170	0.8064
ResNet101	0.9229	0.9000	0.8895	0.8110	0.7994	0.8018	0.7511	0.7578	0.7520
MobileNetV2	0.9467	0.9333	0.9313	0.8640	0.8049	0.8288	0.8925	0.8481	0.8648
MobileNetV3_Small	0.9467	0.9000	0.8952	0.8327	0.7980	0.8131	0.8119	0.8117	0.8076
MobileNetV3_Large	0.9429	0.9333	0.9267	0.8200	0.7994	0.8073	0.8764	8.8392	0.8507
VGG16	0.9667	0.9667	0.9636	0.8651	0.8123	0.8315	0.8702	0.8207	0.8419
VGG19	0.9714	0.9600	0.9624	0.8285	0.7837	0.8032	0.8017	0.8064	0.8002

Figure 12 shows that the learning curve generated during the fine classification of disc multi-view images is not as stable as the learning curve generated for the classification of sagittal disc images using the Mask RCNN model. The training loss curve and validation loss curve continue to decrease with the same trend and are stable within a certain range, but the validation loss curve oscillates up and down within a small range, which is due to the lesser amount of data. When finding the axial images through the center point of discs from the sagittal section, there are missing values exist and some axial images are excluded because they are too far from the center point. In addition, because the samples of medical image data are typically unbalanced, the amount of data belonging to the disc_v1 and disc_v2 disease categories is significantly greater than the amount of data belonging to the other three categories in the generated cone multi-view dataset. To mitigate these issues, we employed some data transformation techniques specifically tailored to medical imaging datasets. These included traditional geometric data augmentation such as small rotations (1–5 degrees) and horizontal flips, along with intensity adjustments to enhance contrast and brightness. Additionally, we used filters, such as Gaussian blur or edge enhancement filters, to simulate different imaging conditions and improve model robustness against variations in image quality. We also introduced random noise into the images to further augment the data. These data augmentation methods were primarily applied to the classes with fewer samples, specifically targeting underrepresented classes to significantly increase their representation, thereby addressing the imbalance in our dataset. These focused augmentation strategies helped stabilize the validation loss curve to

some extent by enhancing the diversity and volume of our training data. However, some oscillations persisted due to the inherent challenges in medical image analysis.

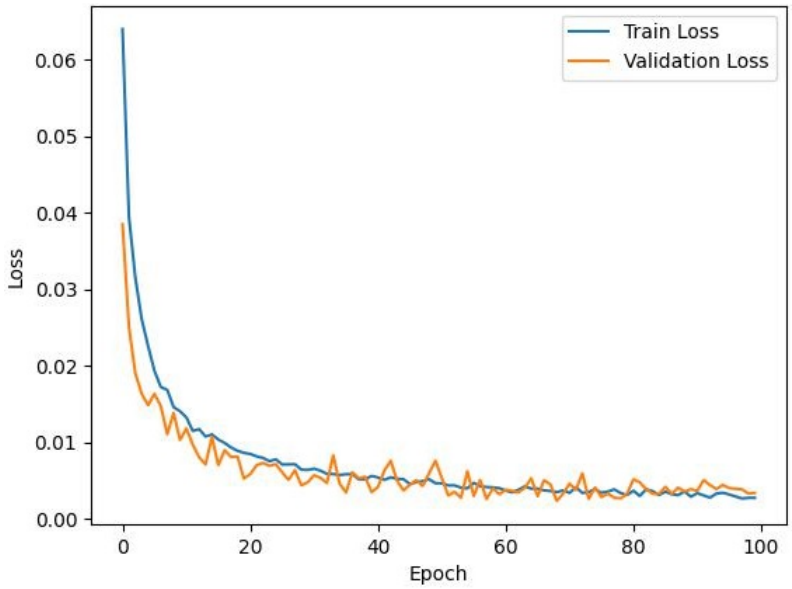

Figure 12. Learning curve of disc multi-view images under VGG16 network.

Therefore, To evaluate the resilience of the model effect, we chose the model VGG16 with the highest F1 score in the experiment for 10-fold cross-validation and output the F1 score performance metric, which has a mean value of 94.43% and a standard deviation of 0.03%. The results are rather consistent.

5.4. Classification Results of Vertebra

We still use nine types of networks to classify vertebrae, such as ResNet, MobileNet, and VGG, and their derivative networks, for feature extraction and data learning. Table 8 shows the three performance metrics of the nine classification networks in the vertebral subclassification experiments. The findings of the experiment indicate that the ResNet50 model has a higher F1 score of 85.27%. The comparison between the F1 scores of each network and their derivative networks reveals that the increase in the number of layers of networks raises the F1 score to a certain extent, but there is a tendency for the F1 score to decrease in the presence of excessively large networks. This is mostly due to the fact that as the number of weight parameters in a model increases, it can better understand the characteristics of the vertebra, but if there are too many parameters, the model will be overfitted. The key to improving the effect lies in balancing the number of parameters.

Table 8. Nine types of network classification effects for vertebra sub-classification.

Model	Precision	Recall	F1 Score
ResNet18	0.6746	0.7292	0.6889
ResNet34	0.6746	0.7291	0.6889
ResNet50	0.8354	0.8354	0.8527
ResNet101	0.74	0.7083	0.7217
MobileNetV2	0.62	0.6042	0.6104
MobileNetV3_Small	0.6491	0.6667	0.6563
MobileNetV3_Large	0.6731	0.625	0.64
VGG16	0.9444	0.75	0.8039
VGG19	0.74	0.7083	0.7217

5.5. Results Summary and Discussion

When using the Mask RCNN model for positioning tasks and rough classification, the highest F1 score obtained is 44.55%. After the classification of the multi-angle view of Disc and vertebra classification based on ResNet Network, the F1 score of the disc is as high as 96.36%, and the F1 score obtained by the vertebral classification has also reached 85.27%. The model classification effect has been greatly improved, which show s the following:

1. Fitting the spine curve to slice the vertebral body and intervertebral disc can ensure that the features of the vertebral body and intervertebral disc are more contained in the region of interest so that they are not easily interfered with by other features. Such an approach has a positive effect on the classification effect.

2. Separating the intervertebral disc and the vertebral body in the network training can obtain a better classification effect.

3. The multi-angle view of the intervertebral disc formed by splicing the axial slice and the sagittal slice of the disc is more advantageous than the separate sagittal or axial slice of the intervertebral disc.

6. Conclusions and Future Work

6.1. Conclusions

The development of an efficient, intelligent, and accurate lumbar spine diagnostic system will be a problem that humans need to solve in the future. This paper proposes using a deep learning framework to detect and classify intervertebral discs and vertebrae in two steps. This method first uses the Mask RCNN model to locate and coarsely classify the vertebral body and intervertebral disc and then puts the vertebra and disc into the ResNet feature extraction network and classification network for training. This article also proposes a method of combining the sagittal plane and the axial plane to enhance the characteristics of the lumbar spine. The model can achieve the effect of the SoTA model in the identification of intervertebral discs and vertebral bodies. This article explains the realization principles of medical images and the data storage structure of their annotations in detail, as well as their preprocessing methods, the realization of masks, and the processing of slices. At the same time, it describes the realization of the model and the process of experimentation, making exploratory attempts for future research. The results of the experiment showed that in the multi-level detection of intervertebral discs, the F1 score reached 96.36%, and the F1 score of the vertebra classification task was 85.27%, which can achieve the effect of medical automatic assistance. The dataset for this experiment comes from real hospital MRI images and has more noisy images, so the model has engineering practicality and robustness.

Furthermore, the integration of this system into existing healthcare workflows could improve the reliability of the diagnostic processes, particularly by connecting the system directly to MRI or CT scanners. This would allow for preliminary diagnoses immediately post-scanning, reducing dependency on manual diagnostics and potentially decreasing the necessity for follow-up visits. Especially beneficial in remote areas lacking sufficient medical expertise, this system could substantially lower misdiagnosis rates and offer cost-effective solutions. Authoritative doctors could further augment the system's accuracy by incorporating their diagnostic assessments back into the database, enhancing the model through continuous learning.

Looking forward, the performance of the system could be improved by employing various advanced deep learning architectures to refine accuracy. Once the single model's accuracy is sufficiently high, the system could transition from semi-automatic to fully automated operations. Ultimately, the future of medical diagnosis models will gradually be transformed into artificial intelligence pre-checks, significantly alleviating the challenges associated with accessing medical care. The system is anticipated to become increasingly pivotal in future research, playing a transformative role in medical diagnostics.

6.2. Research Limitations and Further Work

As discussed previously, the proposed two-stage classification process presents a better performance by synthesizing a multi-angle view of the intervertebral disc, which provides more information for deep learning approaches when classifying lumbar diseases. Although the results look appealing, it still has some limitations. First, the results generated in this research are only based on one dataset. There does exist the probability that the results may shift if the datasets are collected from another hospital, region, or country or even if images are taken by a device from another company. Investigating the performance of the process scheme using more datasets is definitely a meaningful topic and is worthy of further research. Second, the quality of the dataset is also a crucial issue. In this research, the data quality of the training and testing datasets are the same. However, this assumption may not hold for other applications. It is interesting to conduct a sensitivity analysis to know how much the performance of the proposed scheme will shift if the quality of the training and testing datasets are not balanced. This topic could be a good direction for further research. Finally, the results obtained from the AI models discussed in this research are not compared with the performance of radiologists due to limited resources status. It could be interesting and meaningful to further investigate this issue in future studies.

Author Contributions: Conceptualization, K.C. and H.Z.; methodology, K.C.; validation, Z.W.; formal analysis, K.C.; resources, L.Z.; writing—original draft, K.C. and H.Z.; writing—review and editing, L.Z. and H.Z.; visualization, K.C.; supervision, H.Z.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Macao Foundation with grant number MF-24-008-MSB.

Data Availability Statement: The data will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu, A.; Dong, W.; Liu, S.; Cheung, J.P.Y.; Kwan, K.Y.H.; Zeng, X.; Zhang, K.; Sun, Z.; Wang, X.; Cheung, K.M.C.; et al. The prevalence and years lived with disability caused by low back pain in China, 1990 to 2016: Findings from the global burden of disease study 2016. *Pain* **2019**, *160*, 237. [CrossRef] [PubMed]
2. Hartvigsen, J.; Hancock, M.J.; Kongsted, A.; Louw, Q.; Ferreira, M.L.; Genevay, S.; Hoy, D.; Karppinen, J.; Pransky, G.; Sieper, J.; et al. What low back pain is and why we need to pay attention. *Lancet* **2018**, *391*, 2356–2367. [CrossRef] [PubMed]
3. Buchbinder, R.; van Tulder, M.; Öberg, B.; Costa, L.M.; Woolf, A.; Schoene, M.; Croft, P.; Hartvigsen, J.; Cherkin, D.; Foster, N.E.; et al. Low back pain: A call for action. *Lancet* **2018**, *391*, 2384–2388. [CrossRef] [PubMed]
4. Foster, N.E.; Anema, J.R.; Cherkin, D.; Chou, R.; Cohen, S.P.; Gross, D.P.; Ferreira, P.H.; Fritz, J.M.; Koes, B.W.; Peul, W.; et al. Prevention and treatment of low back pain: Evidence, challenges, and promising directions. *Lancet* **2018**, *391*, 2368–2383. [CrossRef] [PubMed]
5. Shoeibi, A.; Ghassemi, N.; Khodatars, M.; Moridian, P.; Khosravi, A.; Zare, A.; Gorriz, J.M.; Chale-Chale, A.H.; Khadem, A.; Rajendra Acharya, U. Automatic diagnosis of schizophrenia and attention deficit hyperactivity disorder in rs-fMRI modality using convolutional autoencoder model and interval type-2 fuzzy regression. *Cogn. Neurodynamics* **2023**, *17*, 1501–1523. [CrossRef] [PubMed]
6. Ghassemi, N.; Shoeibi, A.; Rouhani, M. Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images. *Biomed. Signal Process. Control* **2020**, *57*, 101678. [CrossRef]
7. Natalia, F.; Young, J.C.; Afriliana, N.; Meidia, H.; Yunus, R.E.; Sudirman, S. Automated selection of mid-height intervertebral disc slice in traverse lumbar spine MRI using a combination of deep learning feature and machine learning classifier. *PLoS ONE* **2022**, *17*, e0261659. [CrossRef] [PubMed]
8. Niemeyer, F.; Galbusera, F.; Tao, Y.; Kienle, A.; Beer, M.; Wilke, H.J. A deep learning model for the accurate and reliable classification of disc degeneration based on MRI data. *Investig. Radiol.* **2021**, *56*, 78–85. [CrossRef]
9. Šušteršič, T.; Ranković, V.; Milovanović, V.; Kovačević, V.; Rasulić, L.; Filipović, N. A Deep Learning Model for Automatic Detection and Classification of Disc Herniation in Magnetic Resonance Images. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 6036–6046. [CrossRef]
10. Tong, N.; Gou, S.; Yang, Y.; Liu, B.; Bai, Y.; Liu, J.; Ding, T. Fully Automatic Fine-Grained Grading of Lumbar Intervertebral Disc Degeneration Using Regional Feature Recalibration Network. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 3042–3054. [CrossRef]
11. Xu, Y.; Zheng, S.; Tian, Q.; Kou, Z.; Li, W.; Xie, X.; Wu, X. Deep Learning Model for Grading and Localization of Lumbar Disc Herniation on Magnetic Resonance Imaging. *J. Magn. Reson. Imaging* **2024**. [CrossRef] [PubMed]

12. Pham, D.L.; Xu, C.; Prince, J.L. Current methods in medical image segmentation. *Annu. Rev. Biomed. Eng.* **2000**, *2*, 315–337. [CrossRef] [PubMed]
13. Huang, S.H.; Chu, Y.H.; Lai, S.H.; Novak, C.L. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. *IEEE Trans. Med. Imaging* **2009**, *28*, 1595–1605. [CrossRef] [PubMed]
14. Kadoury, S.; Labelle, H.; Paragios, N. Automatic inference of articulated spine models in CT images using high-order Markov Random Fields. *Med. Image Anal.* **2011**, *15*, 426–437. [CrossRef]
15. Wang, Z.; Zhen, X.; Tay, K.; Osman, S.; Romano, W.; Li, S. Regression segmentation for M^3 spinal images. *IEEE Trans. Med. Imaging* **2014**, *34*, 1640–1648. [CrossRef]
16. Vania, M.; Mureja, D.; Lee, D. Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels. *J. Comput. Des. Eng.* **2019**, *6*, 224–232. [CrossRef]
17. Kolařík, M.; Burget, R.; Uher, V.; Říha, K.; Dutta, M.K. Optimized high resolution 3D dense-U-Net network for brain and spine segmentation. *Appl. Sci.* **2019**, *9*, 404. [CrossRef]
18. Schmidt, S.; Kappes, J.; Bergtholdt, M.; Pekar, V.; Dries, S.; Bystrov, D.; Schnörr, C. Spine detection and labeling using a parts-based graphical model. In *Proceedings of the Biennial International Conference on Information Processing in Medical Imaging, Kerkrade, The Netherlands, 2–6 July 2007;* Springer: Berlin/Heidelberg, Germany, 2007; pp. 122–133.
19. Alomari, R.S.; Corso, J.J.; Chaudhary, V.; Dhillon, G. Lumbar spine disc herniation diagnosis with a joint shape model. In *Computational Methods and Clinical Applications for Spine Imaging*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 87–98.
20. Chen, H.; Shen, C.; Qin, J.; Ni, D.; Shi, L.; Cheng, J.C.; Heng, P.A. Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015;* Springer: Berlin/Heidelberg, Germany, 2015; pp. 515–522.
21. Suzani, A.; Seitel, A.; Liu, Y.; Fels, S.; Rohling, R.N.; Abolmaesumi, P. Fast automatic vertebrae detection and localization in pathological ct scans-a deep learning approach. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015;* Springer: Berlin/Heidelberg, Germany, 2015; pp. 678–686.
22. Bounds, D.G.; Lloyd, P.J.; Mathew, B.G.; Waddell, G. A multilayer perceptron network for the diagnosis of low back pain. In Proceedings of the ICNN, San Diego, CA, USA, 24–27 July 1988; Volume 2, pp. S481–S489.
23. Koompairojn, S.; Hua, K.A.; Bhadrakom, C. Automatic classification system for lumbar spine X-ray images. In Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), Salt Lake City, UT, USA, 22–23 June 2006; pp. 213–218.
24. Ghosh, S.; Raja'S, A.; Chaudhary, V.; Dhillon, G. Computer-aided diagnosis for lumbar mri using heterogeneous classifiers. In Proceedings of the 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Chicago, IL, USA, 30 March–2 April 2011; pp. 1179–1182.
25. Hao, S.; Jiang, J.; Guo, Y.; Li, H. Active learning based intervertebral disk classification combining shape and texture similarities. *Neurocomputing* **2013**, *101*, 252–257. [CrossRef]
26. Oktay, A.B.; Albayrak, N.B.; Akgul, Y.S. Computer aided diagnosis of degenerative intervertebral disc diseases from lumbar MR images. *Comput. Med. Imaging Graph.* **2014**, *38*, 613–619. [CrossRef]
27. Castro-Mateos, I.; Pozo, J.M.; Lazary, A.; Frangi, A.F. 2D segmentation of intervertebral discs and its degree of degeneration from T2-weighted magnetic resonance images. In *Medical Imaging 2014: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Bellingham, WA, USA, 2014; Volume 9035, p. 903517.
28. Jamaludin, A.; Lootus, M.; Kadir, T.; Zisserman, A.; Urban, J.; Battié, M.C.; Fairbank, J.; McCall, I. Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur. Spine J.* **2017**, *26*, 1374–1383. [CrossRef]
29. Shinde, J.; Joshi, Y.; Manthalkar, R. Intervertebral Disc Classification Using Deep Learning Technique. In *Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering, Palladam, India, 16–17 May 2018*; Springer: Berlin/Heidelberg, Germany, 2018, pp. 551–563.
30. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
33. Simonyan, K.; Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*; Computational and Biological Learning Society: Leesburg, VA, USA, 2015; pp. 1–14.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
36. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.

37. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
38. Mustra, M.; Delac, K.; Grgic, M. Overview of the DICOM standard. In Proceedings of the 50th International Symposium ELMAR, Zadar, Croatia, 10–12 September 2008; Volume 1, pp. 39–44.
39. Lu, J.T.; Pedemonte, S.; Bizzo, B.; Doyle, S.; Andriole, K.P.; Michalski, M.H.; Gonzalez, R.G.; Pomerantz, S.R. Deep Spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Palo Alto, CA, USA, 17–18 August 2018; pp. 403–419.
40. Tsai, M.D.; Jou, S.B.; Hsieh, M.S. A new method for lumbar herniated inter-vertebral disc diagnosis based on image analysis of transverse sections. *Comput. Med. Imaging Graph.* **2002**, *26*, 369–380. [CrossRef] [PubMed]
41. Gallucci, M.; Limbucci, N.; Paonessa, A.; Splendiani, A. Degenerative disease of the spine. *Neuroimaging Clin. N. Am.* **2007**, *17*, 87–103. [CrossRef] [PubMed]