




Research Article

End to End Multitask Joint Learning Model for Osteoporosis Classification in CT Images

Kun Zhang ^{1,2,3} **Pengcheng Lin**,¹ **Jing Pan**,⁴ **Peixia Xu**,¹ **Xuechen Qiu**,⁵ **Danny Crookes**,⁶ **Liang Hua** ¹ and **Lin Wang** ⁴

¹School of Electrical Engineering, Nantong University, Nantong, Jiangsu 226001, China

²Nantong Key Laboratory of Intelligent Control and Intelligent Computing, Nantong, Jiangsu 226001, China

³Nantong Key Laboratory of Intelligent Medicine Innovation and Transformation, Nantong, Jiangsu 226001, China

⁴Department of Radiology, Affiliated Hospital 2 of Nantong University, Nantong, Jiangsu 226001, China

⁵College of Mechanical Engineering, Donghua University, Shanghai 201620, China

⁶School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK

Correspondence should be addressed to Liang Hua; hualiang@ntu.edu.cn and Lin Wang; wanglin_nt@126.com

Received 10 December 2022; Revised 23 February 2023; Accepted 1 March 2023; Published 15 March 2023

Academic Editor: Yugen Yi

Copyright © 2023 Kun Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Osteoporosis is a significant global health concern that can be difficult to detect early due to a lack of symptoms. At present, the examination of osteoporosis depends mainly on methods containing dual-energyX-ray, quantitative CT, etc., which are high costs in terms of equipment and human time. Therefore, a more efficient and economical method is urgently needed for diagnosing osteoporosis. With the development of deep learning, automatic diagnosis models for various diseases have been proposed. However, the establishment of these models generally requires images with only lesion areas, and annotating the lesion areas is time-consuming. To address this challenge, we propose a joint learning framework for osteoporosis diagnosis that combines localization, segmentation, and classification to enhance diagnostic accuracy. Our method includes a boundary heat map regression branch for thinning segmentation and a gated convolution module for adjusting context features in the classification module. We also integrate segmentation and classification features and propose a feature fusion module to adjust the weight of different levels of vertebrae. We trained our model on a self-built dataset and achieved an overall accuracy rate of 93.3% for the three label categories (normal, osteopenia, and osteoporosis) in the testing datasets. The area under the curve for the normal category is 0.973; for the osteopenia category, it is 0.965; and for the osteoporosis category, it is 0.985. Our method provides a promising alternative for the diagnosis of osteoporosis at present.

1. Introduction

Osteoporosis (OP) is a disease characterized by impaired bone microstructure and decreased bone mineral density (BMD). With the acceleration of population aging, OP has become an increasingly serious global health problem [1]. Fragile fracture is the most serious complication of OP [2]. OP causes more than 8.9 million brittle fractures each year worldwide [3]. In the US, fragile fractures are more than four times more common than stroke, acute myocardial infarction, and breast cancer [4]. In several developed countries, osteoporotic fractures account for longer

hospitalization time than these diseases according to a meeting of the World Health Organization [5]. By 2025, the number of fragility fractures is expected to increase from 3.5 million in 2010 to 4.5 million, a 28% increase [6]. Therefore, reliable technology for the early detection and prevention of OP is urgently needed.

Currently, although dual-energyX-ray absorptiometry (DXA) is the gold standard for measuring bone mineral density for the diagnosis of OP, it is not widely used as a screening tool for OP owing to its high cost and limited availability of equipment [7]. To overcome these limitations, a variety of osteoporosis screening tools have emerged.

Quantitative ultrasound (QUS) is one of them, which has developed into an alternative method for DXA screening of osteoporosis. Its benefits include being portable and economical; however, it may be unavailable in all primary medical settings [8]. In addition, a variety of clinical risk assessment tools have been developed to predict osteoporosis, including the fracture risk assessment tool (FRAX), the QFracture algorithm, the Garvan Fracture Risk Calculator, and the osteoporosis self-assessment tool [9]. Unfortunately, these tools are based on a combination of known risks to calculate the risk of fracture in patients and have poor efficiency.

Artificial intelligence and machine learning algorithms have recently been used in the diagnosis and prediction of osteoporosis [10]. The existing methods have achieved some success in solving the problem of binary classification (osteoporosis and nonosteoporosis) of which the main purpose is to identify whether the patient has osteoporosis [11]. However, these methods also have some obvious shortcomings: (1) the existing artificial intelligence algorithms treat segmentation and classification as two separate tasks, ignoring the information fusion and complementarity between the two tasks; (2) taking the average of two lumbar cancellous bone mineral density measurements (commonly the first and second lumbar) is widely acknowledged as the best diagnostic criterion for osteoporosis in lumbar QCT [12]. In current models, these data inputs tend to be CT images of a single vertebral body, disregarding the information fusion and complementarity between multiple vertebral images; (3) the problem of class imbalance in the collected data is prevalent due to the lack of standard public datasets; (4) most methods treat osteoporosis as a binary problem, regardless of the urgent need and a strong incentive to turn the binary into a trinomial (osteoporosis, osteopenia, and normal) problem. Although the three classifications are more difficult, osteopenia can bring some predictability to the prevention and treatment of osteoporosis. In this paper, we address the challenges above in the diagnosis of osteoporosis to facilitate the timely detection of the condition and propose an instance-based and class-based multilevel joint learning framework for bone state classification. The innovation of this method lies in the following steps. Firstly, we locate a vertebral body and remove redundant information from the image. Secondly, by constructing the boundary heat map regression auxiliary branch, the vertebral edge is refined, and the segmentation performance is improved on the segmentation branch of the shared encoder. In addition, low-level and high-level features from the segmentation branch and the auxiliary branch, including the shape and boundary of the vertebral body, are fused with feature layers from the diagnostic classifier. Finally, considering the different effects of different vertebral bodies on the classification results of bone state, we design a feature fusion module to adaptively learn feature fusion weights. The proposed method is novel because it solves the challenges of high dimensionality, multimodality, and multiclassification associated with osteoporosis diagnosis, and these challenges have not been resolved in earlier methods. The contributions of the research are as follows:

- (i) A joint learning framework is proposed to segment vertebral bodies from CT images and classify bone states (normal, osteopenia, and osteoporosis)
- (ii) An instance-based and class-based data sampling balancing strategy is introduced to solve the problem of poor model prediction caused by imbalanced data between training datasets
- (iii) A boundary heat map regression branch is proposed, which uses the Gaussian function to do “soft labeling,” accelerating network convergence and improving the performance of vertebral segmentation in joint learning and single-task learning environments
- (iv) The effectiveness of segmentation features in guiding a deep classification network is verified by hierarchically fusing the features of the decoder and classifier related to two segmentation tasks
- (v) A feature fusion module is proposed to adaptively learn the feature weights of vertebrae 1 and 2 and balance the influence of two vertebrae images on classification results

To our knowledge, there are many studies [13–16] on the classification of bone status using vertebral images, but there are few studies on multitask joint learning and detection of bone status based on soft tissue window images at the central level of lumbar 1 and lumbar 2 vertebrae. Experimental results show that multitask joint learning can improve the accuracy of disease classification.

2. Related Works

In this section, we briefly review the related research on bone state classification, categorizing them into three subareas to introduce the current research on the bone state in the medical image, i.e., vertebral positioning, vertebral CT image segmentation, and vertebral medical CT image classification.

2.1. Vertebral Positioning. With the development of deep learning, convolutional neural networks are increasingly used for positioning tasks. However, most of these works describe vertebral recognition as a centroid point detection task. Chen et al. used the advanced features of convolutional neural networks to represent vertebrae from 3D CT volume and eliminated the detection of misplaced centroids based on a random forest classifier [17]. Dong et al. iterated the centroid probability map of a convolutional neural network using a message-passing scheme according to the relationship between the centroids of the vertebrae and used sparse regularization to optimize the localization results to obtain a pixel-level probability of each vertebral centroid [18]. However, it may be more meaningful to directly identify the labels and bounding boxes of vertebrae (rather than the probability map of the centroid point). Zhao et al. proposed a category-consistent self-calibration recognition system to accurately predict the bounding boxes of all vertebrae, improving the discrimination ability of vertebrae categories and the self-awareness of false positive detection

[19]. All of these methods identify the vertebrae from the coronal plane, whereas what we want is to get a small image from the transverse view that only contains the vertebrae.

2.2. Vertebral Segmentation. Recently, machine learning is increasingly used in the recognition and segmentation of vertebral bodies. Michael Kelm et al. used iterative variants of edge-space learning to find the bounding boxes of intervertebral discs and utilized Markov-based random fields and graphical cutting to initialize and guide the segmentation of the vertebrae [20]. Zukić et al. employed the AdaBoost-based Viola–Jones object detection framework to find the bounding boxes of the vertebrae and then split them by expanding the mesh from the center of each vertebra [21]. Chu et al. applied random forest regression to detect the vertebral center and used these to define target regions for the segmentation of the vertebrae with random forest voxel classification [22]. Although these methods can find certain vertebral bodies with specific appearances, they still need to set some parameters empirically and fail to deal with complex pathological cases. However, many recent segmentation methods are based on deep learning, using convolutional neural networks instead of the traditional explicit modeling of spine shape and appearance. For example, Sekuboyina et al. used a multiclass convolutional neural network for pixel labeling, segmented the lumbar spine on a 2D facet slice, and estimated the bounding boxes of the waist region using a simple multilayer perceptron to identify regions of interest in the graph [23]. Janssens et al. depended on two continuous networks to realize this task. First, they used a regression convolutional neural network to estimate the bounding box of the lumbar region and then used a classification convolutional neural network to perform voxel labeling in the bounding box to segment the vertebral body [24]. Mushtaq et al. used ResNet-UNet to semantically segment the lost vertebral body, achieving 0.97 DSC and 0.86 IOU [25].

2.3. Vertebral Medical Image Classification. In the study of establishing the osteoporosis model, Yoo et al. established a support vector machine model using age, height, weight, body mass index, hypertension, hyperlipidemia, and other factors to identify osteoporosis in postmenopausal women. Compared with traditional osteoporosis self-assessment tools, they found that the support vector machine model is more accurate [26]. Pedrassani de Lira et al. established a J48 decision number model to identify osteoporosis through multiple indicators such as age, previous fracture, number of previous fractures, and previous spinal fractures [27]. Taфраouti et al. extracted features from X-ray images and used a support vector machine model to identify osteoporosis, which can well distinguish osteoporosis patients from normal people [28]. Kilic and Hosgormez studied the identification of osteoporosis based on a random subspace method and random forest ensemble model. Jang et al. used a deep learning method to identify osteoporosis [29]. In the internal and external test sets, the area under curve (AUC) of osteoporosis screening was 0.91 (95% confidence interval

(CI), 0.90–0.92) and 0.88 (95% confidence interval (CI), 0.85–0.90), respectively. The experimental results illustrate that the use of chest radiographs based on deep learning models may be used for opportunistic automatic screening of osteoporosis patients in the clinical environment [30]. In the latest study, Xue et al. conducted a study in which they labeled the L1–L4 vertebral body in CT images and divided it into three categories based on bone mineral density: osteoporosis, osteopenia, and normal. The study achieved a high level of accuracy, with a prediction accuracy of 83.4% and a recall rate of 90.0% [31]. Dzierzak and Omiotek have developed a novel method for diagnosing osteoporosis through the use of spine CT imaging and deep convolutional neural networks. To address the issue of a small sample size, they utilized a large dataset to pretrain their model, which resulted in the successful classification of osteoporosis and normal cases. This approach showed promising results for the accurate diagnosis of osteoporosis using CT scans [32]. In these methods, both the traditional machine learning algorithm and the current popular deep learning algorithm use the image containing only the region of interest as the data source. The step-by-step preprocessing process is tedious, time-consuming, and inefficient. Therefore, the integration of positioning, segmentation, and classification into a network should help to improve efficiency, and no research has shown that explicit or implicit features related to the first 3/4 of the vertebral body can be effectively and interpretably used in deep classification networks.

3. Proposed Methods

3.1. Overview. Our proposed method aims to classify vertebral images within a joint framework to enable a more flexible diagnosis of osteoporotic lesions. To achieve this goal, as shown in Figure 1, we propose an instance-based and class-based end-to-end multitask joint learning framework. It mainly has a strategy to solve class imbalance and four deep learning modules, including vertebral positioning module, vertebral segmentation module, cascade feature extraction module combined with gated attention, and feature fusion module. As shown in Figure 2, a new multilayer and multilevel joint learning framework is introduced, which integrates positioning, segmentation, and classification. Firstly, realizing the accurate location of the target lesion (coronal vertebral body), removing the redundant information of the image through the reduction of resolution (from 512×512 to 224×224). Secondly, the boundary heat map auxiliary branch is employed to refine the edge to improve the performance of segmentation; meanwhile, segmentation features are cascaded with the classification features to improve the accuracy of classification. Finally, we propose a feature fusion module, which adaptively assigns feature weights to fuse the features of lumbar L1 and lumbar L2. Different magnitudes of losses in multitask learning tend to bring about negative effects on other tasks when the model tends to fit a certain task; to balance this problem, we use the gradient update method to assign weights to each loss, exploiting neural networks to update the weight parameters.

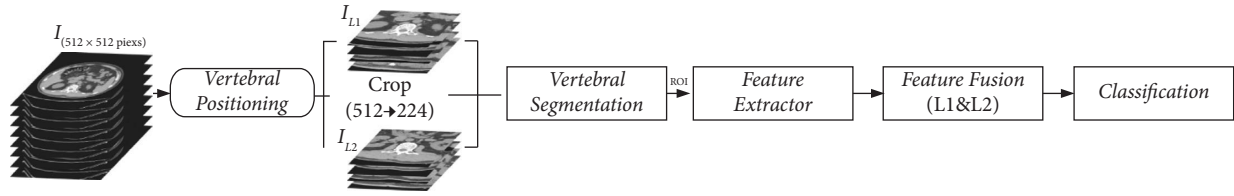


FIGURE 1: Joint framework scheme, including vertebral positioning module and vertebral segmentation module, combined with gated attention cascade feature extraction module and feature fusion module (L1 and L2).

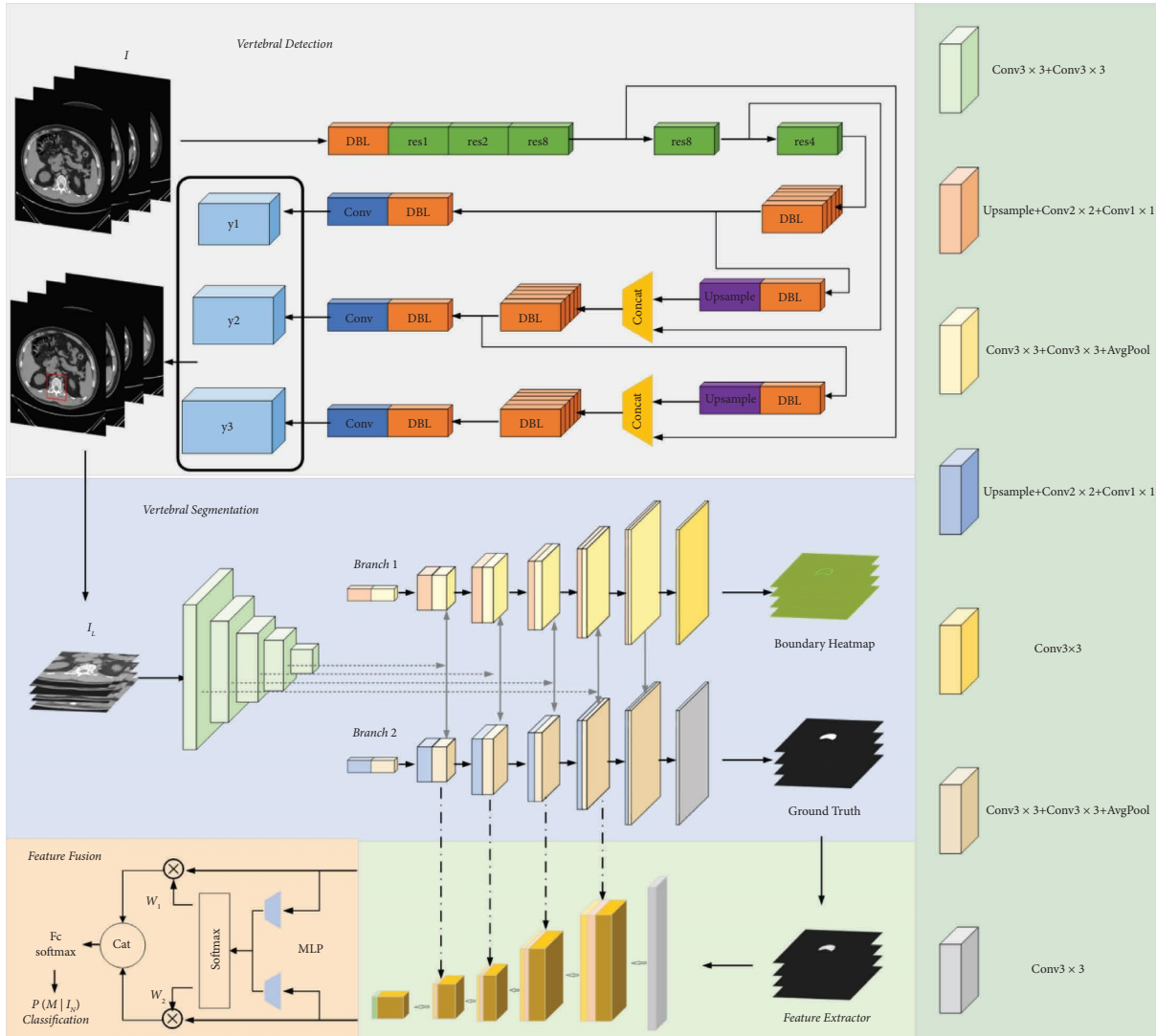


FIGURE 2: Joint framework scheme specific network architecture, including (i) the CT image is sent to YOLOv3 for vertebral positioning; (ii) then, the segmentation module is used to segment the region of interest of the vertebral body, and the feature maps of different scales of the decoding layer are cascaded with the features learned by the ResNet-based convolutional feature extractor, and the key features are obtained by modeling the context features through gated attention; (iii) finally, the features of L1 and L2 are fused by the feature fusion module, and the CT image is classified.

3.2. *Instance and Class-Based Sampling Methods.* In the actual clinical scene, the data collected by image acquisition will be unbalanced owing to the inherent difficulty of collecting labels of rare diseases or other unusual cases. Therefore, when training on extremely unbalanced data, the model may have a high probability of being affected by the

number of different categories, resulting in the underfitting of some categories which may be ignored. At present, the methods to solve the data imbalance include data resampling [33], adaptive loss function [34], and curriculum learning [35]. Inspired by the paper [36, 37], methods are introduced to solve the problem of extreme imbalance of our category

images. It combines unbalanced (instance-based) and balanced (class-based) sampling of data, where we extend the method to our three-category practical problems.

We define the training set as $D = \{(x_i, y_i), i = 1, 2, \dots, N\}$, where x_i is the sample, y_i is the sample category. Assuming that for multiclassification problems with K categories, each category has M_k samples, and N represents the total number of samples, where $\sum_{k=1}^K M_k = N$, the general sampling strategies can be described as

$$p_j = \frac{M_j^n}{\sum_{k=1}^K M_k^n}, \quad (1)$$

where p_j is the probability of sampling from the j th category. If we set $n = 0$, the probability of sampling from each category is equal to $1/K$. This is the class-based sampling method.

If we set $n = 1$, then it is equivalent to selecting the sample by the proportion of a category of samples to all samples, which is instance-based sampling. Here, we introduce a mixed sampling method based on instance and class, which is suitable for data imbalance. We denote the training dataset and sampling strategy by the symbol (D, S) . Instance-based sampling and class-based sampling are represented by S_I and S_C , respectively, so this mixture can be described as

$$\begin{aligned} \hat{x} &= \lambda x_I + (1 - \lambda)x_C, \\ \hat{y} &= \lambda y_I + (1 - \lambda)y_C, \end{aligned} \quad (2)$$

where $\lambda \sim \text{beta}(\alpha, \beta)$, $\alpha > 0, \beta > 0, \lambda \in [0, 1]$, $(x_I, y_I) \in (D, S_I)$, $(x_C, y_C) \in (D, S_C)$. \hat{x} and \hat{y} represent random convex combinations of data and label inputs. Here, we set $\beta = 1$. As shown in Figure 3, as α grows, examples from minority classes are combined with a greater weight to avoid overfitting of minority classes. Here, we set $\alpha = 0.1$ to induce a more balanced distribution of training samples by creating synthetic data points around spatial regions where minority classes provide fewer data density.

3.3. Vertebral Positioning Module Based on YOLOv3. The basic step of vertebral CT image classification is to extract robust features from CT images, given W and H of the original images are 512 pixels. To remove redundant features, we use the YOLOv3 [38] to locate the vertebral body in the image with size $512 \times 512 \times 3$ as input to YOLOv3. The image feature is extracted by DarkNet-53, and then the target classification and position regression are performed on the acquired feature map with the help of the FPNs (feature pyramid networks) structure.

In this study, we will obtain the position of the prediction box in the original image p_x, p_y, p_w, p_h , in YOLOv3, a set of anchor frames is composed of nine initial frames of different sizes. Assuming that the center coordinates, width, and height of an anchor frame are expressed as a_x, a_y, a_w, a_h ,

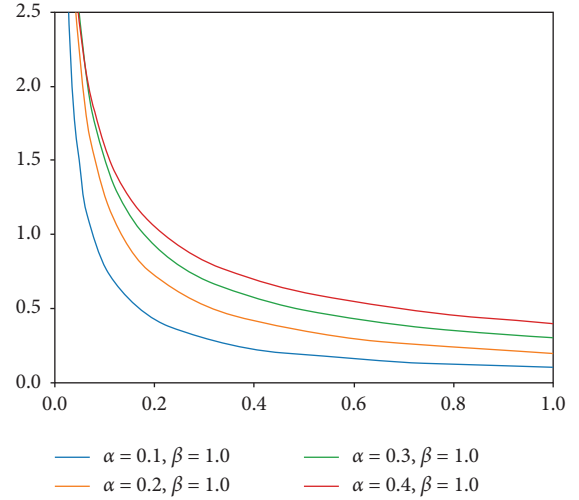


FIGURE 3: Beta($\alpha, 1$) distribution for varying α .

p_x, p_y, p_w, p_h can be obtained by reverse calculation of the regression parameter t_x, t_y, t_w, t_h by the output network. Details of the calculation formula are as follows:

$$\begin{aligned} p_x &= \sigma(t_x) + a_x, \\ p_y &= \sigma(t_y) + a_y, \\ p_w &= a_w e^{t_w}, \\ p_h &= a_h e^{t_h}, \end{aligned} \quad (3)$$

where $\sigma(\cdot)$ represents the sigmoid transformation of the variable, aiming at controlling the offset of the center point between 0 and 1.

The main purpose of employing YOLOv3 is to obtain the center coordinates p_x and p_y of the prediction box and utilize this position as the center cutting position of the vertebral body to obtain a 224×224 image containing the complete vertebral body as the input of the subsequent convolution module. In this way, we can remove tens of thousands of useless features and improve the efficiency of the model.

3.4. Boundary Regression Auxiliary Branches. We suggest dividing the segmentation task into two tasks: vertebral segmentation and contour determination. Thus, our network is mainly composed of a weight-sharing encoder and two decoders composed of the segmentation branch and boundary regression branch. In the encoder, we improve the original U-Net [39] by applying residual blocks to replace the original two effective convolutions. In the decoder stage, we cascade the penultimate features from the boundary regression branch with the penultimate features of the segmentation branch, helping the network to better perceive and refine the vertebral contour. Since vertebrae in CT images may show up hyperosteoegeny or other conditions, it is necessary to reconstruct edges by constructing auxiliary tasks, which provide more explicit and implicit topological

priors for the coding layer and enable them to assist with the segmentation branches to obtain more accurate target masks.

The problem of boundary inaccuracy is rooted in the similarity of information in the corresponding receptive field of pixels. When similar features belong to the interior or exterior of the segmented region, this similarity will be advantageous, inversely similar information lies in the segmented boundary will undoubtedly increase the uncertainty of the edge. In terms of the boundary regression auxiliary branch in the segmentation module, we propose to divide the edge based on the region and graph from the whole image, combining it with the spatial proximity and pixel value similarity. In this paper, the accurate boundary of vertebral segmentation should be the inner boundary. We combine the convolutional neural network with the level set, taking the segmentation result obtained by the neural network as the prior knowledge of level set segmentation; then we construct a gray level constraint term on the original level set function and improve the edge indicator function to deal with uneven intensity in the image.

3.4.1. Improve the Edge Indicator Function. Getreuer [40] proposed the famous Chan–Vese (CV) model in 2001. This method uses a region-based segmentation strategy to divide the image into two homogeneous regions, the inner and outer regions, using active contoured lines to find the image to be segmented and the original image with the minimum difference to minimize the energy function.

Given the input image $I(x, y)$, the energy function based on the CV is shown as follows:

$$\begin{aligned} E(C, C_1, C_2) = & \mu \int_{\Omega} g \delta(\phi) |\nabla \phi| dx dy \\ & + v \int_{\Omega} g H(-\phi) dx dy \\ & + \lambda_1 \int_{\Omega_1} |I - C_1|^2 dx dy \\ & + \lambda_2 \int_{\Omega_2} |I - C_2|^2 dx dy, \end{aligned} \quad (4)$$

where C_1 and C_2 describe the average gray levels of equivalent parts inside and outside the contour, respectively, Ω_1 and Ω_2 represent the inner and outer regions of the contour, λ_1 , λ_2 , μ , v are constants, $g = (1/1 + |\nabla G(x, y, \sigma) * I(x, y)|)$ is the edge indicator function which can be used to prevent the curve from exceeding the target area, G is the Gaussian calculation sub, σ is the standard deviation, and δ and H represent Dirac and Heaviside functions, respectively. The position of contour C and unknowns $C_1(\phi)$ and $C_2(\phi)$ are finally obtained through optimization formula (4).

The evolution of the CV model is constrained by global gray-level information. However, most images, especially medical images, have uneven intensity. To solve this problem, we improve the function g and construct gray-level information constraint terms to constrain the evolution

direction. Bilateral filtering is a method that combines the spatial proximity of images with the similarity of pixel values. Based on Gaussian filtering, bilateral filtering introduces the gray value of pixels for the local weighted average. When smoothing the speckle noise of images, bilateral filtering can better maintain the edge features.

In the first step, the Gaussian function $G_{sr}(x, y, \sigma)$ is used to construct bilateral filters to obtain smooth images:

$$\begin{aligned} G_{sr}(x, y, \sigma) &= G_{\sigma_s} * G_{\sigma_r}, \\ G_{\sigma_s} &= e^{-(x-k)^2 + (y-l)^2 / 2\sigma_s^2}, \\ G_{\sigma_r} &= e^{-\|I(x, y) - f(k, l)\|^2 / 2\sigma_r^2}. \end{aligned} \quad (5)$$

Image $I(x, y)$ is filtered using bilateral filter operator $g(x, y) = G_{sr}(x, y, \sigma) \cdot I(x, y)$, where σ_r is the standard deviation used to control the smoothness, i, j, k, l are the weight coefficients.

In the second step, the optimal threshold T is calculated based on the filtered image using the adaptive threshold principle. The maximized interclass variance value of T is shown in the following equation:

$$\gamma^2 = w_0 \times w_1 \times (u_0 - u_1), \quad (6)$$

where w_0 represents the ratio of pixels in the target area to the image, u_0 represents the corresponding average gray level, w_1 is the proportion of background pixels, and u_1 is the average gray level of background pixels. Then, the new edge indicator function g_r can be described as

$$g_r = \frac{1}{1 + \gamma^2 |\nabla G_{sr}(x, y, \sigma) * I(x, y)|}. \quad (7)$$

3.4.2. Auxiliary Branch. We advocate the segmentation results of convolutional neural networks as prior knowledge, namely, the initial contour of the level set, and the curve contour evolved through the level set is used to guide the neural network to optimize toward the edge of the vertebral body.

The specific expression of the gray level constraint Q is described as

$$\begin{aligned} Q &= \gamma \left[\frac{1 + \Gamma}{2} - \frac{1 - \Gamma}{2} \right] H(\phi), \\ \Gamma &= \begin{cases} -1, & I \in (I_{\text{low}}, I_{\text{high}}), \\ 1, & I \notin (I_{\text{low}}, I_{\text{high}}), \end{cases} \end{aligned} \quad (8)$$

$$I_{\text{low}} = \eta - w \cdot \sigma,$$

$$I_{\text{high}} = \eta + w \cdot \sigma,$$

where I_{high} is the upper limit of the vertebral gray value obtained by using the convolutional neural network model,

I_{low} is the lower limit of vertebral gray value, σ is the average of vertebral gray value, η is the variance of vertebral gray value, and w is a constant.

The function of the gray level information constraint term is to make the level set curve evolve inside the vertebral body to approximate the inner edge contour. When the gray value of the pixel is within the upper and lower limits of the initial vertebral gray value, the energy value of the point is negative, otherwise positive. The edge result obtained by the neural network is used to replace x and y on the initial contour plane. Gradient descent is used to minimize the energy function, and the formula form of the final evolution equation after adding the gray constraint function is shown as follows:

$$\frac{\partial \phi}{\partial t} = \delta(\phi) \begin{bmatrix} \mu \operatorname{div} \left(g_T \frac{\nabla \phi}{|\nabla \phi|} \right) - g_T v \\ -\lambda_1 [I(x, y) - C_1]^2 + \gamma \\ \left[\frac{1 + \Gamma}{2} - \frac{1 - \Gamma}{2} \right] + \lambda_2 [I(x, y) - C_2]^2 \end{bmatrix},$$

$$C_1(\phi) = \frac{\int_{\Omega} I(x, y) H(\phi) dx dy}{\int_{\Omega} H(\phi) dx dy},$$

$$C_2(\phi) = \frac{\int_{\Omega} I(x, y) [1 - H(\phi)] dx dy}{\int_{\Omega} [1 - H(\phi)] dx dy},$$

$$\phi_0 = \phi(0, I(x, y)).$$

(9)

In the label aspect of the auxiliary branch, we use the Canny operator to detect the edge of the binary image label. Canny is built on a two-dimensional convolution. To improve the calculation speed of the Canny operator, two-dimensional convolution can be decomposed into one-dimensional filters, and then a convolution operation with the image $A(x, y)$ is carried out, respectively: $E_x = (\partial G / \partial x) \cdot A(x, y)$, $E_y = (\partial G / \partial y) \cdot A(x, y)$. Then, the gradient amplitude $A(x, y)$ and gradient $a(x, y)$ direction can be expressed as

$$A(x, y) = \sqrt{E_x^2(x, y) + E_y^2(x, y)},$$

$$a(x, y) = \arctan \frac{E_y(x, y)}{E_x(x, y)}. \quad (10)$$

The size of the Gaussian window is adjusted by changing the standard deviation σ of the Gaussian function, that is $A(x, y) = \max(\sqrt{E_x^2 + E_y^2})$. We first apply nonmaximum suppression, and then segment images through the dual-threshold method. When the gradient of some pixel is greater than the limit threshold, it will be considered as an edge pixel.

Then, we construct a soft label heat map in the form of Heatsum based on the processed images:

$$\text{Heatsum}(G(x_1, y_1, \sigma), G(x_2, y_2, \sigma)) = 1 - (1 - G(x_1, y_1, \sigma))(1 - G(x_2, y_2, \sigma)),$$

$$G_{bd} = \text{Gaussheat}(\partial G),$$

$$= \text{Heatsum}(G(x_1, y_1, \sigma), \dots, G(x_n, y_n, \sigma)), \forall G(x_n, y_n, \sigma) \in \partial G, \quad (11)$$

where \odot represents the Hadamard product; it is noted that G_{bd} is normalized between $[0, 1]$.

Here, the boundary regression branch is utilized to refine the segmented edges. We treat this branch as a regression task through mean square error rather than a whole work consisting of a boundary segmentation task together with the segmentation branch.

3.5. Cascading Classification Module. In the classification module, we use ResNet-101 as a basic feature extractor. ResNet [41] is a traditional deep convolutional neural network where the residual structure is used in the shallow network. The corresponding structure is illustrated in Figure 4(b). By adding the input value x with the output unit, the residual gains better performance in convergence after the operation of ReLU active. These steps can be

approximated as an identical mapping of equal input and output, which effectively solves the problems of network learning ability decline, gradient disappearance, and gradient explosion when the number of convolutional neural network layers increases.

Inspired by the gating attention [42] and residual structure, we designed a gating residual module as shown in Figure 4 to replace the first convolution module in ResNet-101 from conv2_x to conv5_x. The specific network parameters can be found in Figure 5. The gated residual model can be described as follows.

Assuming that $x \in \mathbb{R}^{C \times H \times W}$ is the activation feature of the convolutional neural network, where H and W are the height and width of the image, and C is the number of channels of the image, in general, the gating attention performs the following transformation.

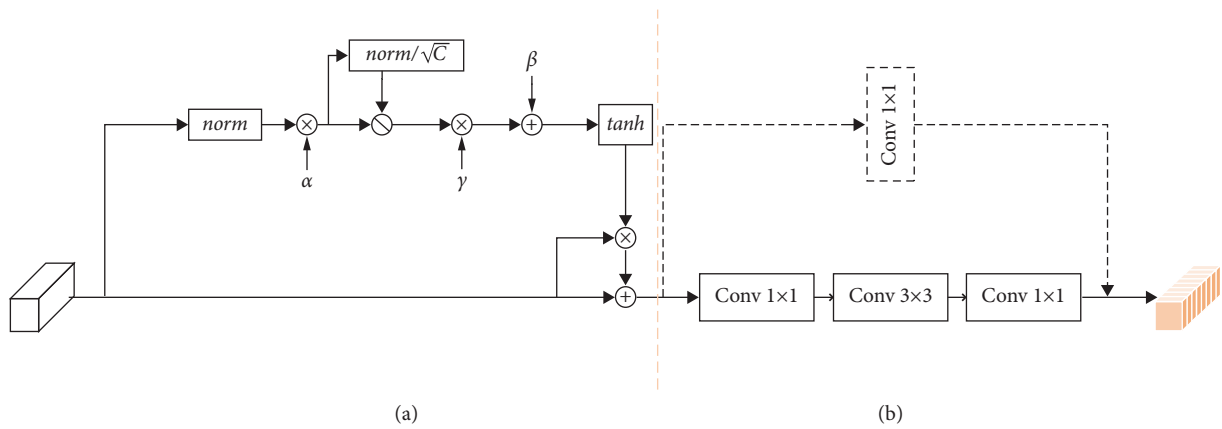


FIGURE 4: Gated residual module. (a) Gct layer. (b) Residual layer.

Layer Name	Output Size	Architecture
Conv1	112×112	$7 \times 7, 64, \text{stride } 2$
Conv 2_x	56×56	$3 \times 3, \text{maxpool, stride } 2$
		$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 1 + \begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 2$
		$1 \times 1, 64, \text{stride } 1$
Conv 3_x	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 1 + \begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$
		$1 \times 1, 512, \text{stride } 1$
Conv 4_x	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 1 + \begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 22$
		$1 \times 1, 1024, \text{stride } 1$
Conv 5_x	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 1 + \begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 2$
		$1 \times 1, 2048, \text{stride } 1$
	1×1	Average pool, 2048-d Fc, softmax

FIGURE 5: Feature extraction module network structure diagram. The convolution block framed in red is replaced by a gating residual module.

$$\hat{x} = F(x|\alpha, \gamma, \beta), \alpha, \gamma, \beta \in \mathbb{R}^C. \quad (12)$$

Among them, a , β , and γ are trainable parameters. The embedding weight a is mainly responsible for adjusting the embedding output, and the gating weight γ and the bias weight β are responsible for adjusting the gating activation.

They determine the behavior of gated attention in each channel.

For the specific process, assuming the given embedding weight as $\alpha = [\alpha_1, \alpha_2 \dots, \alpha_c]$, modules can be defined as

$$s_c = \alpha_c \|x_c\|_2$$

$$= \alpha_c \left\{ \left[\sum_{i=1}^H \sum_{j=1}^W (x_c^{i,j})^2 \right] + \epsilon \right\}^{1/2}, \quad (13)$$

$$\hat{s}_c = \frac{\sqrt{C}s_c}{\|S\|_2}$$

$$= \frac{\sqrt{C}s_c}{\left[\left(\sum_{c=1}^C s_c^2 \right) + n \right]^{1/2}}, \quad (14)$$

where ϵ is a small constant, which is mainly used to avoid the derivation of zeros. Equation (14) is used to normalize channels, and n represents a small constant. \sqrt{C} is used for normalization the ratio of s_c , preventing the condition of small s_c when C is too large, α_c is a trainable parameter used for controlling the weight of each channel. When α_c is close to 0, the channel will not participate in channel normalization.

Then, we suppose the selection weight $\gamma = [\gamma_1, \gamma_2 \dots, \gamma_c]$ and the gating offset $\beta = [\beta_1, \beta_2 \dots, \beta_c]$, the gating function can be depicted as follows:

$$\hat{x}_c = x_c [1 + \tanh(\gamma_c \wedge s_c + \beta_c)]. \quad (15)$$

Each primitive channel x_c is adapted by the corresponding gate, γ and β are trainable weights and deviations which is used to control the activation of the gate. Finally, $\hat{x}_c = [\hat{x}_1, \hat{x}_2 \dots, \hat{x}_c]$ will be entered into the residual

module to obtain the feature map $y = [y_1, y_2 \dots, y_c]$ of the gating attention after the convolution operation. Supposing the feature map concatenated from the segmentation module $S = [S_1, S_2 \dots, S_c]$, we can perform the following operations on the classification network feature $y = [y_1, y_2 \dots, y_c]$ and the segmentation module feature to obtain the final feature map \hat{y}_c .

$$\hat{y}_c = \text{Conv}_{1 \times 1}(\text{Concat}(y_c, S_c))_{i=1,2,\dots,c}. \quad (16)$$

Two 1×2048 -dimensional feature vectors of vertebrae can be obtained by flattening the feature map.

3.6. Feature Fusion Module. As mentioned above, the detection of bone status is based on the average of lumbar L1 and lumbar L2. To explain the different effects of different lumbar vertebrae on classification, we learn W_1 and W_2 adaptively for each vertebra, which satisfies $W_1 + W_2 = 1$; W_1 and W_2 represent the fusion weights, respectively.

$$X_{\text{fuse}} = \text{Concat}(W_1 \times X_1, W_2 \times X_2). \quad (17)$$

Specifically, we calculate W_1 and W_2 ($W_1 + W_2 = 1$) by $F_{\text{fuse}}(X_1)$ and $F_{\text{fuse}}(X_2)$, respectively, where F represents the perception of two layers, that is, two fully connected layers. The following *softmax* layer can be used to eliminate the influence of different feature dimensions. After gaining the feature X_{fuse} , the prediction of bone state $P(M|I_N)$ can be given by the fully connected layer and softmax function.

$$P(M|I_N) = \text{softmax}(fc(X_{\text{fuse}}, \text{num} - \text{classes})). \quad (18)$$

3.7. Cascading Classification Models. To balance the impact of different dimensions of multiple tasks in the training process we introduce the trade-off parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 to balance these four tasks. The total loss function of multitask learning can be defined as

$$L_{\text{mul}} = \lambda_1 L_{\text{IOC}} + \lambda_2 L_{\text{cla}} + \lambda_3 L_{\text{conf}} + \lambda_4 L_{\text{seg}} + \lambda_5 L_{\text{seg}} + \lambda_6 L_{\text{cla}}$$

$$= \lambda_1 L_{\text{IOC}}(p_i^I, t) + \lambda_2 L_{\text{sobj}}(p_i^{c1}, q_{\text{cla1}}) + \lambda_3 L_{\text{obj}}(p_i^{c2}, q) + \lambda_4 L_{\text{dice}}(S(p_i^s), G)$$

$$+ \lambda_5 L_{\text{mse}}(p_i^b, G_{bd}^n) + \lambda_6 L_{\text{crossentropy}}(p_i^{c3}, q_{\text{cla2}}), \quad (19)$$

where $p_i^I, p_i^{c2}, p_i^s, p_i^b, p_i^{c3}$, respectively, represent the predicted results of the positioning branch, category branch, confidence branch, and segmentation branch of the positioning module for a given input image, the boundary heatmap regression branch, and the classification network. S represents the Sigmoid function, t represents the prediction box result, and q_{cla1} is the result of the category in the positioning module. q represents the probability that a vertebral body exists, G_{bd}^n represents the normalized result of G_{bd} , and q_{cla2} is the expected result of the classification network.

4. Experimental Results

4.1. Dataset and Preprocessing. To assess the effectiveness and benefit of the joint learning framework in bone state classification, we conducted experiments in a dataset obtained from the Nantong First People's Hospital from May 2021 to May 2022, consisting of CT images of 1048 routine-dose cases. All images were collected by Ingenuity Core 128 CT (Philips Health Care, Holland), the tube voltage was 120 kV, the inpatient tube current

modulation technique was used, and the iDose 4 was used to reconstruct the cross-sectional image of the mediastinal window (standard B standard reconstruction algorithm). The reconstruction layer thickness and layer interval were both 2 mm. The longitudinal window images of the lumbar 1 and lumbar 2 center planes of each subject were selected for BMD measurement and deep learning model construction. The QCT pro4 software (Mindways, CA, USA) was used to set the same size of the region of interest (ROI) in the central cancellous bone area of the lumbar 1 and lumbar 2 vertebral bodies, avoiding the cortical bone and the visible vascular area. The software automatically calculated the BMD values of the lumbar 1 and lumbar 2 vertebral bodies and used their mean values as the BMD values of the individual subjects (BMD individuals). According to the standard recommended by the “expert consensus on imaging and bone mineral density diagnosis of osteoporosis” BMD individuals $> 120 \text{ mg/cm}^3$ are normal bone mass, $80 \text{ mg/cm}^3 \leq \text{BMD individuals} \leq 120 \text{ mg/cm}^3$ are osteopenia, and BMD individuals $< 80 \text{ mg/cm}^3$ are osteoporosis.

We divide the dataset into training data (50%), validation data (10%), and test data (40%); the class distribution of training, validation, and testing datasets is shown in Figure 6. These three datasets do not have any overlapping images, and the CT images of each category in the three datasets are placed in strict proportions. Then, all images are resized to 512×512 and each image is normalized from $[0, 255]$ to $[0, 1]$ before being fed into the network.

To increase the amount of training data and improve the generalization ability and robustness of the model, we enhance the image data employing flipping, rotating, and scaling on the basis of the original data balancing strategy based on an instance and actual class.

4.2. Implementation of Framework. To implement the joint learning framework, we implemented the model based on Python 3.6.12, using the PyTorch framework and two NVIDIA GeForce 3090Ti GPUs. We apply the SGD optimizer to train the joint learning framework for 300 epochs with a learning rate of $(10e - 1 - 10e - 5)$ and add six adaptive parameters to the SGD optimizer to weigh the loss of multitask learning.

4.3. Measurements and Baselines

4.3.1. Measurements. Based on previous work [49–52], accuracy, sensitivity, specificity, and *F1*-score were used to evaluate the performance of classification. The accuracy rate is the ratio of the number of samples correctly classified by the classifier to the total number of samples. The sensitivity reflects the proportion of positive cases correctly judged by the classifier to the total positive samples. The specificity indicates the proportion of negative cases correctly judged by the classifier to the total negative samples. *F1*-score is the sum of accuracy and sensitivity. In this paper, the three-category problem is transformed into a two-category

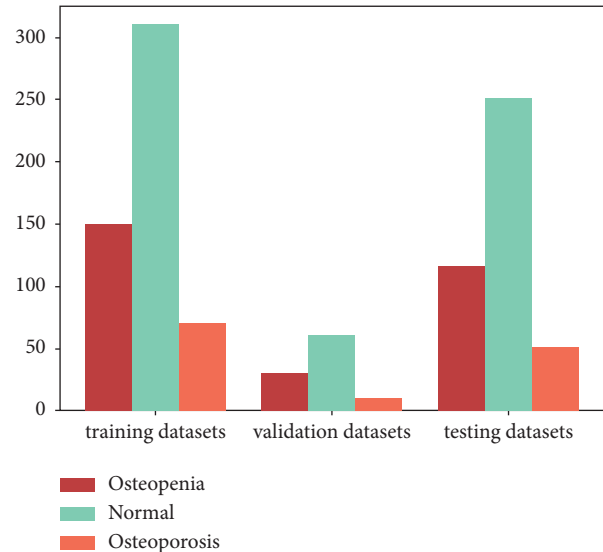


FIGURE 6: Data distribution of datasets.

problem to evaluate; that is, the category studied at this time is a positive sample and the other categories are negative samples.

- (i) Accuracy: $\text{Acc} = (\text{TP} + \text{FN} / \text{TP} + \text{TN} + \text{FP} + \text{FN})$
- (ii) Sensitivity: $\text{Pre} = (\text{TP} / \text{TP} + \text{FN})$
- (iii) Specificity: $\text{Spe} = (\text{TN} / \text{FP} + \text{TN})$
- (iv) *F1*-score: $\text{F1} = (2 \times \text{P} \times \text{R} / \text{P} + \text{R})$, $\text{P} = (\text{TP} / \text{TP} + \text{FP})$, $\text{R} = (\text{TP} / \text{TP} + \text{FN})$

P denotes the model prediction and *T* denotes the true label. Positive samples are predicted as positive samples (true positive, TP), positive samples are predicted as negative samples (false negative, FN), negative samples are predicted as positive samples (false positive, FP), and negative samples are predicted as negative samples (true negative, TN).

Based on previous works [53–55], we use the intersection over union (IOU) and dice coefficient (Dice) to evaluate the effectiveness of our model segmentation task and use the average precision (AP) to evaluate the effectiveness of the positioning task.

- (i) Intersection over union: $\text{IOU} = (\text{TP} / \text{TP} + \text{FP} + \text{FN})$
- (ii) Dice coefficient: $\text{Dice} = (2\text{TP} / 2\text{TP} + \text{FP} + \text{FN})$

4.3.2. Baselines. To demonstrate the performance of our federated framework model, we compared our work with popular machine learning and deep learning methods, including AlexNet [43], VGG-19 [44], GoogLeNet [45], ResNet [41], DenseNet [46], ShuffleNet [47], and EfficientNet [48].

4.4. Results. We use ten-fold cross-validation to calculate the average results and show the performance of the joint framework in Table 1. We set the learning rate of $10e - 1 - 10e - 5$ to evaluate the classification performance of the joint framework in different situations. We used normal

TABLE 1: Comparison of joint framework performance at different learning rates.

Learning rate		Accuracy	Sensitivity	Specificity	F1-score
0.1	Normal	0.876	0.910	0.813	0.862
	Osteopenia	0.816	0.818	0.810	0.865
	Osteoporosis	0.935	0.951	0.823	0.963
0.01	Normal	0.936	0.955	0.923	0.925
	Osteopenia	0.898	0.905	0.880	0.927
	Osteoporosis	0.962	0.983	0.829	0.978
0.001	Normal	0.971	0.964	0.976	0.964
	Osteopenia	0.933	0.970	0.836	0.954
	Osteoporosis	0.957	0.962	0.922	0.975
0.0001	Normal	0.880	0.964	0.825	0.866
	Osteopenia	0.811	0.848	0.716	0.866
	Osteoporosis	0.921	0.921	0.922	0.953
0.00001	Normal	0.900	0.941	0.873	0.882
	Osteopenia	0.864	0.868	0.853	0.902
	Osteoporosis	0.964	0.981	0.843	0.980

The bold value indicates that this is the best model results.

(osteopenia and osteoporosis) as a positive sample and other categories as negative samples, achieving an accuracy of 0.971, a sensitivity of 0.964, a specificity of 0.976, and an $F1$ -score of 0.964. We achieved 0.933 in accuracy, 0.970 in sensitivity, 0.836 in specificity, and 0.954 $F1$ -score when osteopenia was used as a positive sample and other categories (normal and osteoporosis) as a negative sample. When we used osteoporosis as a positive sample and other categories (normal and osteopenia) as negative samples, we achieved an accuracy of 0.957, a sensitivity of 0.962, a specificity of 0.922, and an $F1$ -score of 0.975. The best performance is obtained by the learning rate of $10e-3$, indicating that the classification problem of bone state CT images can be effectively solved by adjusting the hyperparameters.

In addition, we compare the best results of joint learning with the most advanced baselines. The comparison results are reported in Table 2, where the best comparable performance is represented in bold. For the input images of other classification methods, we use CT images (512×512) generated by labels manually drawn by physicians that contain only regions of interest. To better intuitively compare the classification performance of the model, we use the confusion matrix for visual analysis. As shown in Figure 7, joint learning in dealing with the task of identifying low-dose achieves good performance with only 5 cases misclassified as normal, 2 cases misclassified as osteoporosis, and 8 cases misclassified as low doses; in the task of identifying osteoporosis, only 3 cases were misclassified as low dose. This result fully indicates the nonexistence of overfitting and underfitting states; this result further illustrates that there is no bias to a certain category which increases accuracy results.

The histogram of accuracy and $F1$ -score can be found in Figure 8. Intuitively, the accuracy rate has increased. Compared with the highest accuracy rate among advanced baseline methods, the accuracy rate of joint learning has increased by 6.2% in the osteopenia category, 3.3% in the normal category, and 0.1% in the osteoporosis category. Notably, when compared to the overall accuracy of advanced

baseline methods, the overall accuracy of joint learning was improved by 3.8% which proved the effectiveness of joint learning strategies once again.

4.5. Further Discussion

4.5.1. Roc Curve. To better demonstrate the classification ability of our proposed joint learning framework, we use the operating characteristic curve (ROC) and the area under curve (AUC) of receivers as further evaluation indicators. Taking the experimental results with a learning rate of 0.01 as an example, we draw the ROC curves of three categories in Figure 9, AUC for each category is also depicted in the figure. It can be found that the AUC in the osteopenia state is 0.965, the AUC value in the Normal state is 0.973, and the AUC value in the osteoporosis state is 0.985. These values prove the effectiveness of joint learning in bone CT image classification tasks.

4.5.2. Training Convergence. For model training, we use the accuracy and loss curve and the training process to imply the training trend of accuracy and model cost. The accuracy and loss curves of the joint learning framework with a learning rate of 0.01 are shown in Figure 10, which reflects that the model’s performance achieved satisfactory results at the 150th epoch and became stable. These two curves show the convergence of the model and assess its stability in bone CT image classification. In addition, the total training time of the joint learning framework on our dataset is about 10 hours, and each epoch takes 2 minutes. In short, training convergence and time reveal the computational efficiency of our network.

4.5.3. Model Visualization. We further use gradient weighted class activation mapping (Grad-CAM) to visualize the decision information of the feature extraction module. Figure 11 shows that the feature extraction modules for different categories (normal, osteopenia, and osteoporosis) focus on different regions, and the model automatically focuses on the corresponding regions. Compared with the

TABLE 2: Comparison with the state-of-the-art baselines on dataset.

Models		Accuracy	Sensitivity	Specificity	F1-score
AlexNet (2012) [43]	Normal	0.868	0.785	0.785	0.878
	Osteopenia	0.799	0.793	0.793	0.687
	Osteoporosis	0.931	0.882	0.882	0.756
VGG-19 (2014) [44]	Normal	0.856	0.994	0.765	0.847
	Osteopenia	0.756	0.795	0.655	0.825
	Osteoporosis	0.900	0.894	0.941	0.939
GoogLeNet (2015) [45]	Normal	0.899	0.976	0.849	0.886
	Osteopenia	0.811	0.871	0.655	0.869
	Osteoporosis	0.911	0.902	0.980	0.947
ResNet-50 (2016) [41]	Normal	0.911	0.874	0.936	0.888
	Osteopenia	0.868	0.894	0.802	0.907
	Osteoporosis	0.956	0.995	0.986	0.976
ResNet-101 (2016) [41]	Normal	0.938	0.958	0.924	0.925
	Osteopenia	0.871	0.917	0.750	0.911
	Osteoporosis	0.933	0.940	0.882	0.961
DenseNet-121 (2017) [46]	Normal	0.897	0.982	0.840	0.884
	Osteopenia	0.849	0.841	0.871	0.889
	Osteoporosis	0.952	0.967	0.843	0.972
ShuffleNet (2018) [47]	Normal	0.926	0.970	0.896	0.913
	Osteopenia	0.856	0.911	0.716	0.902
	Osteoporosis	0.931	0.924	0.980	0.959
EfficientNet (2019) [48]	Normal	0.926	0.976	0.892	0.913
	Osteopenia	0.871	0.904	0.784	0.910
	Osteoporosis	0.945	0.943	0.961	0.968
Joint framework (ours)	Normal	0.971	0.964	0.976	0.964
	Osteopenia	0.933	0.970	0.836	0.954
	Osteoporosis	0.957	0.962	0.922	0.975

The bold value indicates that this is the best model results.

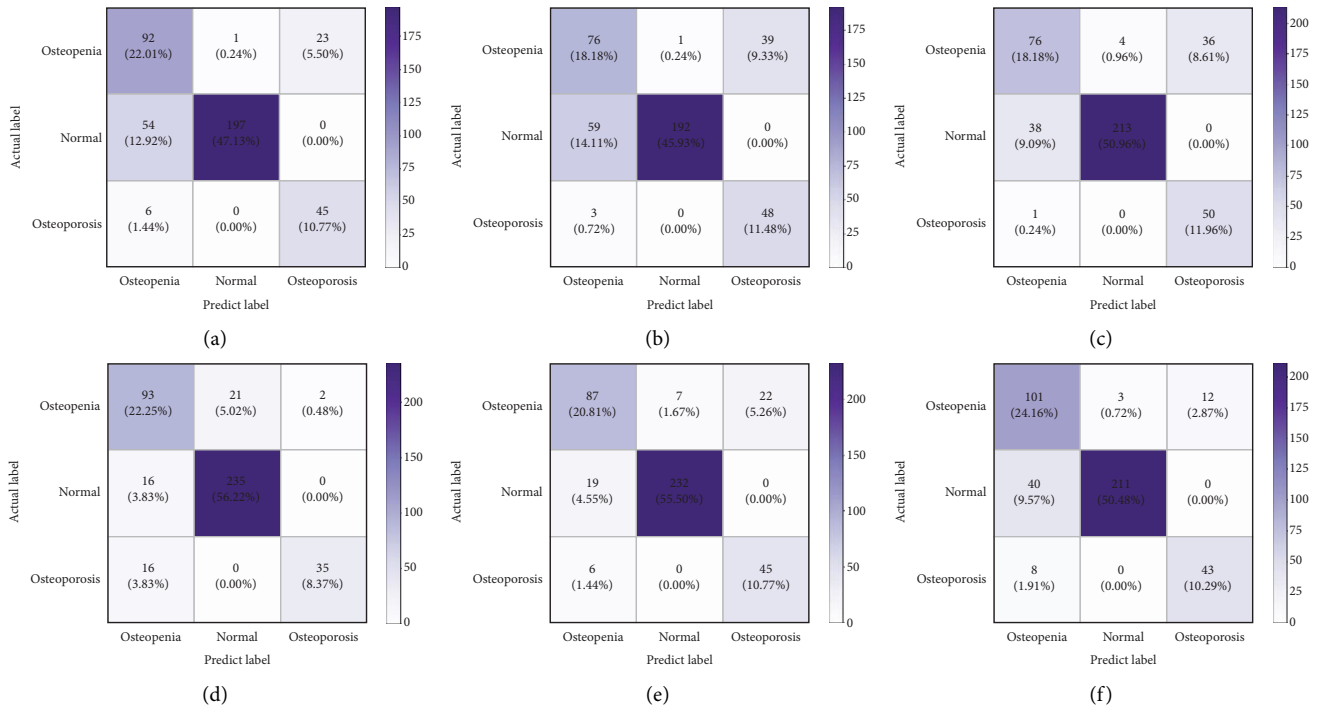


FIGURE 7: Continued.

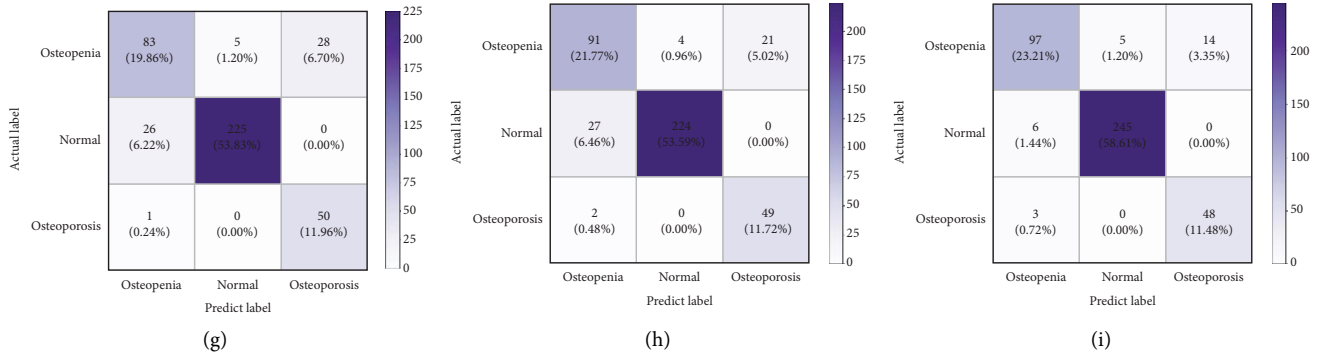


FIGURE 7: (a)–(h) The confusion matrix for each baseline method. (i) The confusion matrix for this method. (a) AlexNet. (b) VGG-19. (c) GoogLeNet. (d) ResNet-50. (e) ResNet-101. (f) DenseNet-121. (g) ShuffleNet. (h) EfficientNet. (i) Joint framework.

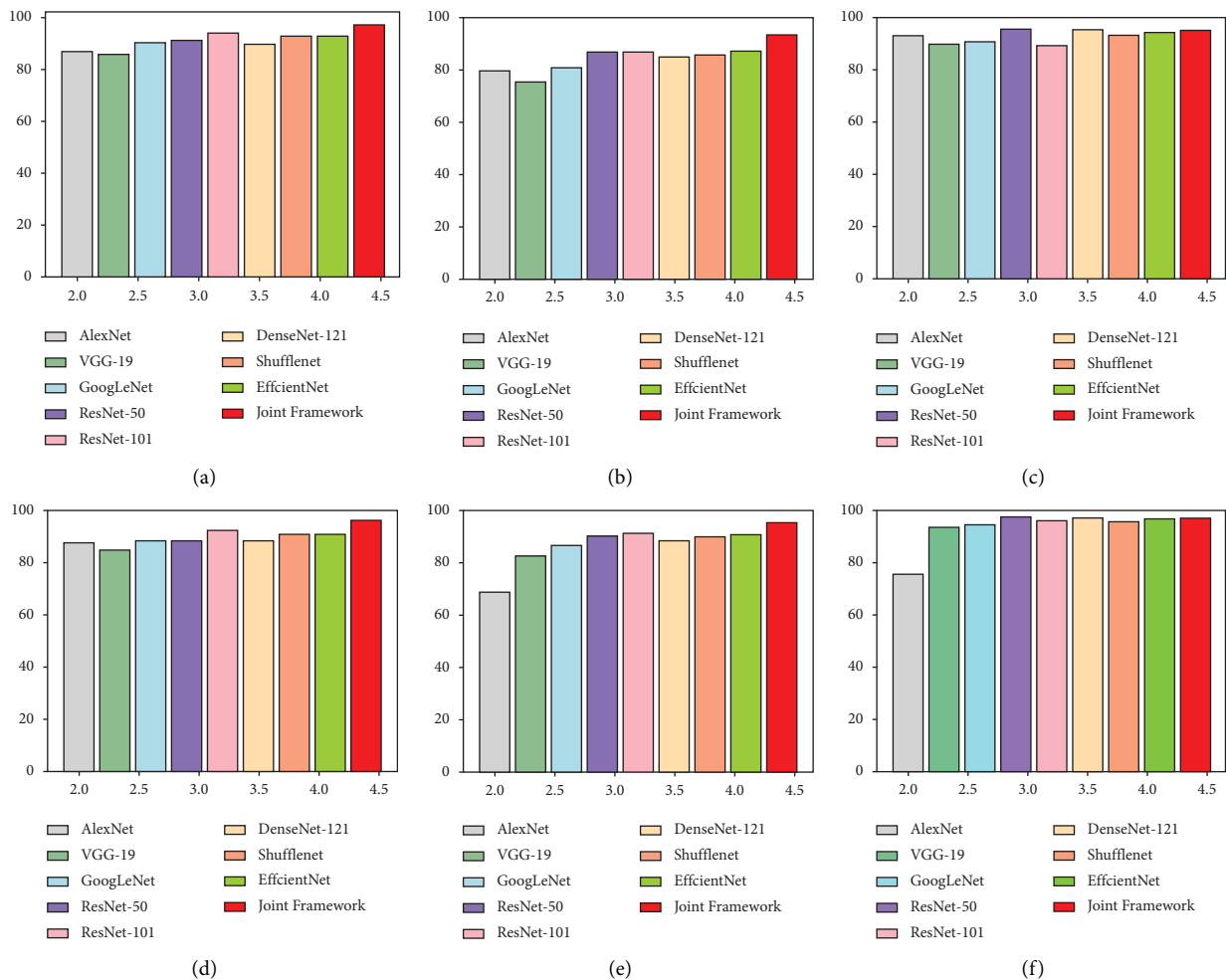
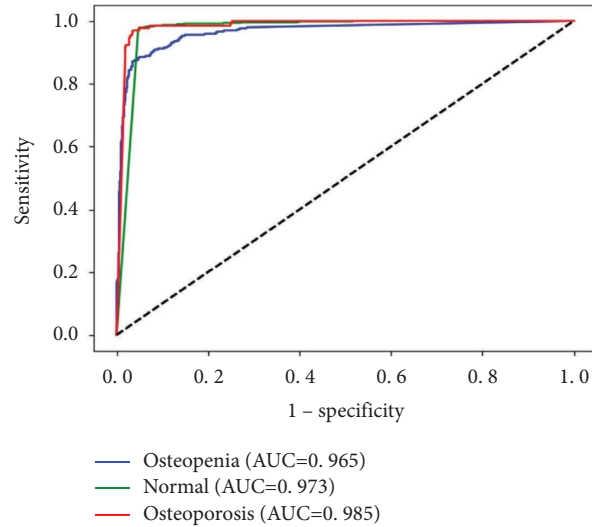
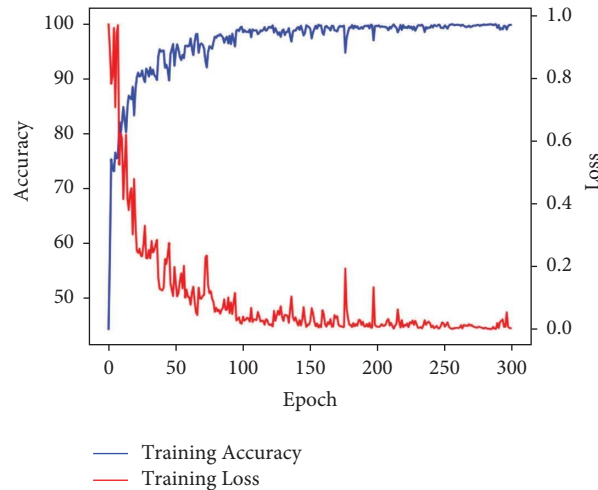


FIGURE 8: The classification performance comparison of each baseline method. (a) Normal-accuracy. (b) Osteopenia-accuracy. (c) Osteoporosis-accuracy. (d) Normal-F1-score. (e) Osteopenia-F1-score. (f) Osteoporosis-F1-score.

correctly classified decision information, we also list some cases of misclassification in Figure 12. The focus area of the wrong case has changed significantly compared with the correct case in Figure 11, which may be used as an explanation for the neural network decision error.

Meanwhile, we calculated that the AP value of all testing datasets in the positioning task is 95%, the average IOU in the segmentation task is 0.972 ± 0.125 , and the average Dice is 0.983 ± 0.036 , which shows that we have good efficiency in selecting features in the

FIGURE 9: The ROC curve for learning rate $10e-3$.FIGURE 10: Accuracy and loss curve for learning rate $10e-3$.

positioning and segmentation tasks, but in some cases, these features have no good effect on classification.

4.5.4. Ablation Experiments. In this section, we conduct an ablation study (learning rate is $10e-3$) of our method to prove the effective impact of segmentation feature and classification feature layered fusion (LF), gated convolution (GC) module, and feature fusion module (FF).

We use the three modules separately and combine them randomly and calculate the overall accuracy of each experiment to evaluate whether the model is improved. The quantitative result can be found in Table 3. In Figure 13, it can be clearly seen that the accuracy of the model has been

greatly improved. When we calculate without using the method of three modules; it is unfortunate to find that the accuracy of the model is only 82.1%. However, when we perform a hierarchical fusion of segmentation features and classification features, the overall accuracy rate rises to 85.6%, an increase of 3.5%. When we use the gated convolution module, we find that the accuracy rate has increased by 2.8%. When we use feature fusion of vertebral bodies at different levels, the overall accuracy rate has increased by 3.3%. When we select any two of them, we find that the overall accuracy rate has increased by 4%, 8.1%, and 10.5%, respectively. The seven additional experiments prove the feasibility and effectiveness of our proposed modular methods in improving classification accuracy.

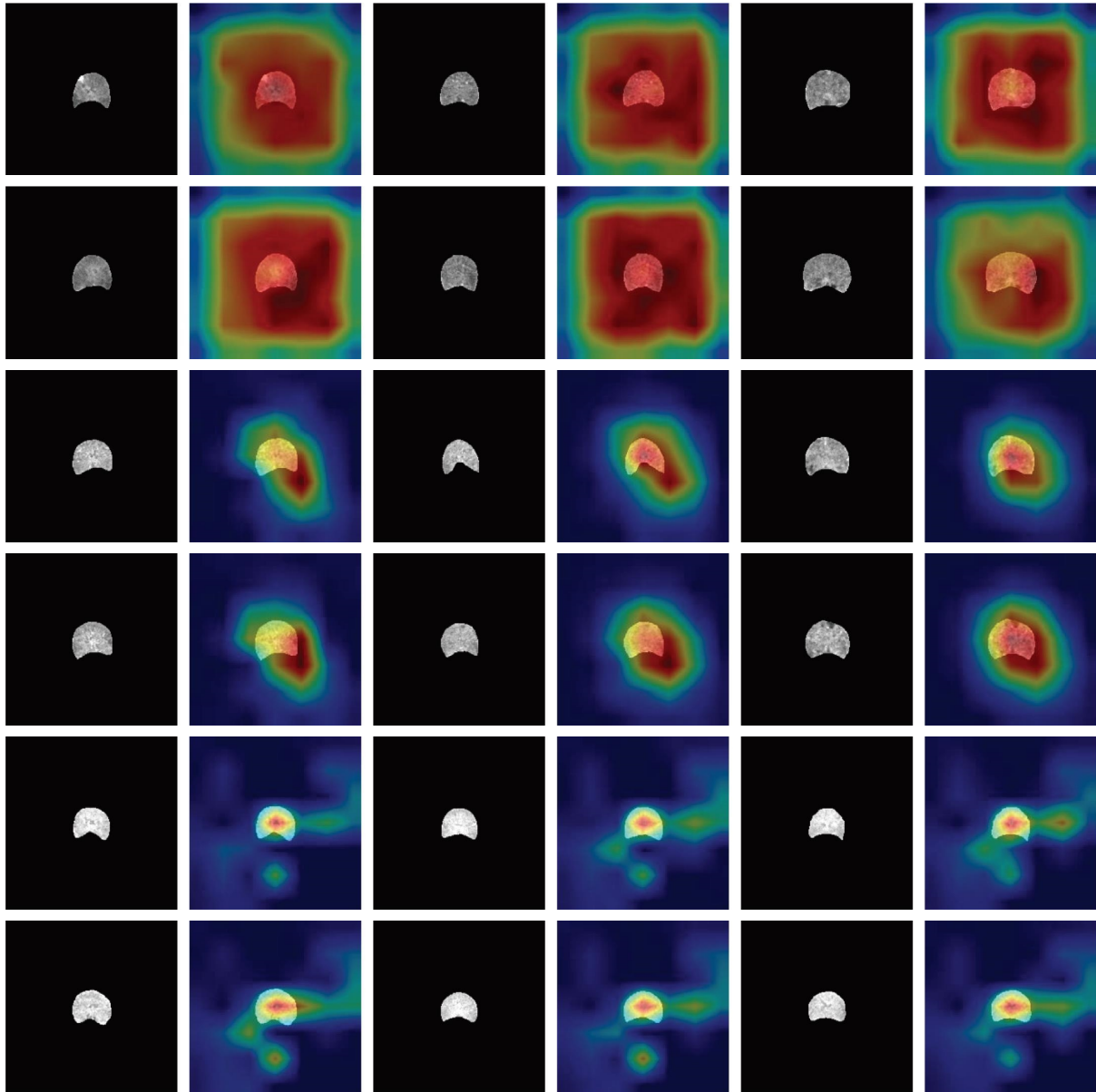


FIGURE 11: Grad-CAM visualization of 9 cases. It can be seen that different categories of networks have different emphases, which can be used as an explanation of neural networks. The first line of each two lines represents the L1 vertebrae, and the second line represents the corresponding L2 vertebrae. The first two lines represent osteoporosis cases, the middle two lines represent osteopenia cases, and the last two lines represent normal cases.

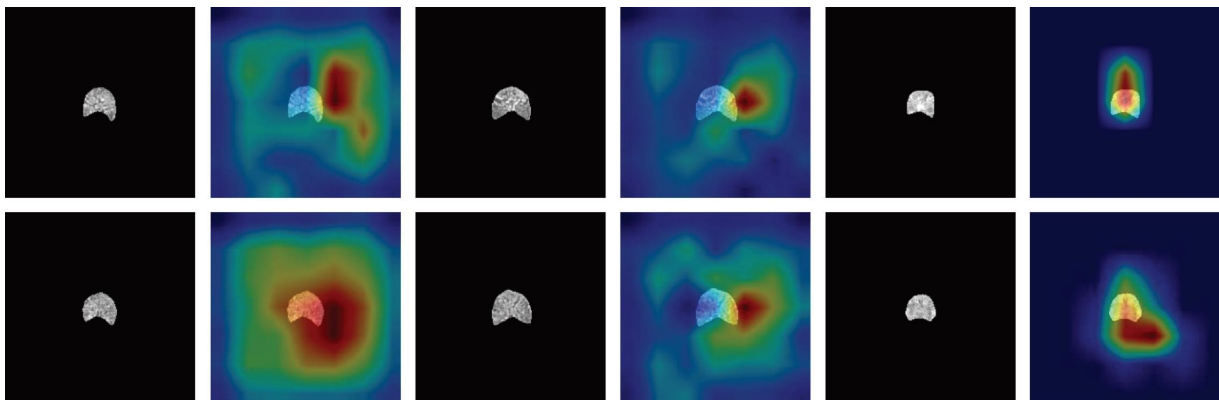


FIGURE 12: Grad-CAM visualization of 3 cases. From left to right are osteoporosis, osteopenia, and normal cases. The first line represents the L1 vertebral body, and the second line represents the L2 vertebral body.

TABLE 3: Ablation experiments in a joint learning framework.

LF	GC	FF	Accuracy	Sensitivity	Specificity	F1-score
		Normal	0.873	0.743	0.960	0.824
		Osteopenia	0.825	0.921	0.578	0.884
		Osteoporosis	0.943	0.978	0.686	0.967
✓		Normal	0.907	0.784	0.988	0.870
		Osteopenia	0.859	0.957	0.603	0.907
		Osteoporosis	0.947	0.970	0.784	0.970
	✓	Normal	0.889	0.802	0.948	0.853
		Osteopenia	0.849	0.921	0.664	0.898
		Osteoporosis	0.959	0.983	0.784	0.977
		Normal	0.907	0.892	0.916	0.884
		Osteopenia	0.854	0.900	0.733	0.899
		Osteoporosis	0.947	0.965	0.824	0.969
✓	✓	Normal	0.904	0.850	0.940	0.877
		Osteopenia	0.861	0.924	0.698	0.906
		Osteoporosis	0.956	0.972	0.843	0.975
	✓	Normal	0.935	0.982	0.904	0.924
		Osteopenia	0.904	0.901	0.914	0.934
		Osteoporosis	0.964	0.978	0.963	0.979
		Normal	0.966	0.946	0.980	0.958
✓	✓	Osteopenia	0.928	0.980	0.793	0.952
		Osteoporosis	0.956	0.956	0.961	0.972

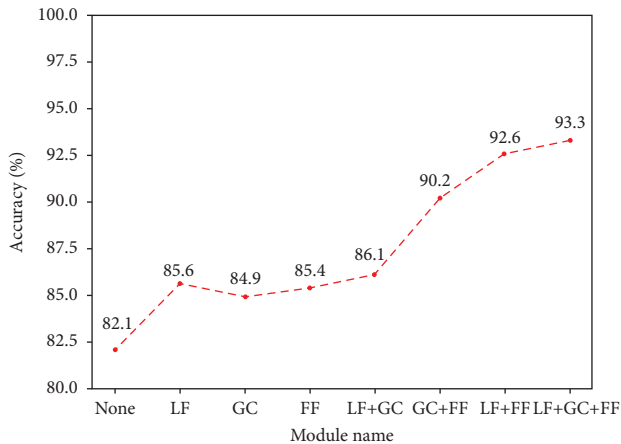


FIGURE 13: The overall accuracy line chart of the ablation experiments.

5. Conclusion

Machine learning can help a great deal in accurately identifying osteoporosis from CT images. In this study, we propose a joint learning framework for bone state detection, where we integrate positioning, segmentation, and classification into an end-to-end multitask joint learning framework. The framework processes from the original input to the final output, increasing the overall fit of the model. The accuracy of classification has been improved by modular task fusion, global feature association, and fusion of different vertebral features. We used a CT image database containing three categories of vertebrae to evaluate this method. A large number of experiments confirm this method improves the overall accuracy from 82.1% to 93.3%, which shows the effectiveness of joint learning in bone state image classification and contributes to solving the problem of clinical diagnosis of osteoporosis.

Data Availability

The data used to support the study are included within the article.

Ethical Approval

This retrospective study was approved by the Ethics Committee of Nantong First People’s Hospital (No.: 2021KT028), who waived the need for informed consent. The study protocol was implemented according to the Good Clinical Practice guidelines defined by the Helsinki Declaration and the International Conference on Harmonisation (ICH).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Nantong Basic Science Research and Social People’s Livelihood Science and Technology Program (MSZ2022009 and MS22021032), Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX21_1449), and Joint Project of Industry-University-Research of Jiangsu Province (BY2022224).

References

- [1] M. Lorentzon and S. R. Cummings, “Osteoporosis: the evolution of a diagnosis,” *Journal of Internal Medicine*, vol. 277, no. 6, pp. 650–661, 2015.
- [2] P. R. Ebeling, H. H. Nguyen, J. Aleksova, A. J. Vincent, P. Wong, and F. Milat, “Secondary osteoporosis,” *Endocrine Reviews*, vol. 43, no. 2, pp. 240–313, 2022.

- [3] The Health Investigators, “Total hip arthroplasty or hemiarthroplasty for hip fracture,” *New England Journal of Medicine*, vol. 381, no. 23, pp. 2199–2208, 2019.
- [4] H. Gharib, E. Papini, R. Paschke et al., “American association of clinical endocrinologists, associazione medici endocrinologi, and European thyroid association medical guidelines for clinical practice for the diagnosis and management of thyroid nodules: executive summary of recommendations,” *Journal of Endocrinological Investigation*, vol. 33, no. 5, pp. 287–291, 2010.
- [5] A. Piccoli, F. Cannata, R. Strollo et al., “Sclerostin regulation, microarchitecture, and advanced glycation end-products in the bone of elderly women with type 2 diabetes,” *Journal of Bone and Mineral Research*, vol. 35, no. 12, pp. 2415–2422, 2020.
- [6] C. Beaudart, L. Lengelé, V. Leclercq et al., “Symptomatic efficacy of pharmacological treatments for knee osteoarthritis: a systematic review and a network meta-analysis with a 6-month time horizon,” *Drugs*, vol. 80, no. 18, pp. 1947–1959, 2020.
- [7] J. S. Kimball, J. P. Johnson, and D. A. Carlson, “Oxidative stress and osteoporosis,” *Journal of Bone and Joint Surgery*, vol. 103, no. 15, pp. 1451–1461, 2021.
- [8] J. J. Carey, P. Chih-Hsing Wu, and D. Bergin, “Risk assessment tools for osteoporosis and fractures in 2022,” *Best Practice & Research Clinical Rheumatology*, vol. 36, no. 3, Article ID 101775, 2022.
- [9] S. Chavda, B. Chavda, and R. Dube, “Osteoporosis screening and fracture risk assessment tool: its scope and role in general clinical practice,” *Cureus*, vol. 14, no. 7, Article ID e26518, 2022.
- [10] N. Yamamoto, S. Sukegawa, A. Kitamura et al., “Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates,” *Biomolecules*, vol. 10, no. 11, p. 1534, 2020.
- [11] R. Jang, J. H. Choi, N. Kim, J. S. Chang, P. W. Yoon, and C. H. Kim, “Prediction of osteoporosis from simple hip radiography using deep learning algorithm,” *Scientific Reports*, vol. 11, no. 1, pp. 19997–19999, 2021.
- [12] X. Cheng, K. Zhao, X. Zha et al., “Opportunistic screening using low-dose CT and the prevalence of osteoporosis in China: a nationwide, multicenter study,” *Journal of Bone and Mineral Research*, vol. 36, no. 3, pp. 427–435, 2021.
- [13] B. Zhang, K. Yu, Z. Ning et al., “Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: a multicenter retrospective cohort study,” *Bone*, vol. 140, Article ID 115561, 2020.
- [14] N. Teclé, J. Teitel, M. R. Morris, N. Sani, D. Mitten, and W. C. Hammert, “Convolutional neural network for second metacarpal radiographic osteoporosis screening,” *Journal of Hand Surgery*, vol. 45, no. 3, pp. 175–181, 2020.
- [15] A. S. Areeckal, N. Jayasheelan, J. Kamath, S. Zawadynski, M. Kocher, and S. David, “Early diagnosis of osteoporosis using radiogrammetry and texture analysis from hand and wrist radiographs in Indian population,” *Osteoporosis International*, vol. 29, no. 3, pp. 665–673, 2018.
- [16] J. Smets, E. Shevroja, T. Hugel, W. D. Leslie, and D. Hans, “Machine learning solutions for osteoporosis—a review,” *Journal of Bone and Mineral Research*, vol. 36, no. 5, pp. 833–851, 2021.
- [17] H. Chen, C. Shen, J. Qin et al., “Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks,” *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 515–522, Springer, Cham, Switzerland, 2015.
- [18] Y. Dong, T. Xiong, D. Xu et al., “Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization,” *International Conference on Information Processing in Medical Imaging*, pp. 633–644, Springer, Cham, Switzerland, 2017.
- [19] S. Zhao, X. Wu, B. Chen, and S. Li, “Automatic vertebrae recognition from arbitrary spine MRI images by a category-consistent self-calibration detection framework,” *Medical Image Analysis*, vol. 67, Article ID 101826, 2021.
- [20] B. Michael Kelm, M. Wels, S. Kevin Zhou et al., “Spine detection in CT and MR using iterated marginal space learning,” *Medical Image Analysis*, vol. 17, no. 8, pp. 1283–1292, 2013.
- [21] D. Zukić, A. Vlasák, J. Egger, D. Hořínek, C. Nimsky, and A. Kolb, “Robust detection and segmentation for diagnosis of vertebral diseases using routine MR images,” *Computer Graphics Forum*, vol. 33, no. 6, pp. 190–204, 2014.
- [22] C. Chu, D. L. Belavy, G. Armbrrecht, M. Bansmann, D. Felsenberg, and G. Zheng, “Fully automatic localization and segmentation of 3D vertebral bodies from CT/MR images via a learning-based method,” *PLoS One*, vol. 10, no. 11, Article ID e0143327, 2015.
- [23] A. Sekuboyina, K. Jan, J. S. Kirschke, B. H. Menze, and V. Alexander, “Attention-driven deep learning for pathological spine segmentation,” *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, pp. 108–119, Springer, Cham, Switzerland, 2017.
- [24] R. Janssens, G. Zeng, and G. Zheng, “Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks,” in *Proceedings of the 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 893–897, IEEE, Washington, DC, USA, April 2018.
- [25] M. Mushtaq, M. U. Akram, N. S. Alghamdi, J. Fatima, and R. F. Masood, “Localization and edge-based segmentation of lumbar spine vertebrae to identify the deformities using deep learning models,” *Sensors*, vol. 22, no. 4, p. 1547, 2022.
- [26] T. K. Yoo, S. K. Kim, and D. W. Kim, “Risk prediction of femoral neck osteoporosis using machine learning and convolutional methods,” *International Work-Conference on Artificial Neural Networks*, pp. 181–188, Springer, Berlin, Heidelberg, 2013.
- [27] C. Pedrassani de Lira, L. L. Toniazco de Abreu, A. C. Veiga Silva et al., “Use of data mining to predict the risk factors associated with osteoporosis and osteopenia in women,” *CIN: Computers, Informatics, Nursing*, vol. 34, no. 8, pp. 369–375, 2016.
- [28] A. Taфраouti, M. El Hassouni, H. Toumi, E. Lespessailles, and R. Jennane, “Osteoporosis diagnosis using fractal analysis and support vector machine,” in *Proceedings of the 2014 10th International Conference on Signal-Image Technology and Internet-Based Systems*, pp. 73–77, IEEE, Marrakech, Morocco, November 2014.
- [29] N. Kilic and E. Hosgormez, “Automatic estimation of osteoporotic fracture cases by using ensemble learning approaches,” *Journal of Medical Systems*, vol. 40, no. 3, pp. 61–10, 2016.
- [30] M. Jang, M. Kim, S. J. Bae, S. H. Lee, J. M. Koh, and N. Kim, “Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset,” *Journal of Bone and Mineral Research*, vol. 37, no. 2, pp. 369–377, 2022.

- [31] L. Xue, Y. Hou, S. Wang et al., "A dual-selective channel attention network for osteoporosis prediction in computed tomography images of lumbar spine," *Acadlore Transactions on AI and Machine Learning*, vol. 1, no. 1, pp. 30–39, 2022.
- [32] R. Dzierżak and Z. Omiotek, "Application of deep convolutional neural networks in the diagnosis of osteoporosis," *Sensors*, vol. 22, no. 21, p. 8189, 2022.
- [33] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.
- [34] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, Washington, DC, USA, April 2019.
- [35] A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff et al., "Medical-based deep curriculum learning for improved fracture classification," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 694–702, Springer, Cham, Switzerland, 2019.
- [36] A. Galdran, G. Carneiro, A. Miguel, and G. Ballester, "Balanced-mixup for highly imbalanced medical image classification," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 323–333, Springer, Cham, Switzerland, 2021.
- [37] K. Lei, M. Mardani, J. M. Pauly, and S. S. Vasanawala, "Wasserstein GANs for MR imaging: from paired to unpaired training," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 105–115, 2021.
- [38] M. Tan, R. Pang, V. Quoc, and Le. Efficientdet, "Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790, Salt Lake City, UT, USA, June 2020.
- [39] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 60–48, 2019.
- [40] P. Getreuer, "Chan-veze segmentation," *Image Processing On Line*, vol. 2, pp. 214–224, 2012.
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [42] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11794–11803, Salt Lake City, UT, USA, June 2020.
- [43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, <https://arxiv.org/abs/1409.1556>.
- [45] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, Honolulu, HI, USA, July 2015.
- [46] G. Huang, Z. Liu, L. van der Maaten, and Q. Kilian, "Weinberger. Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.
- [47] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, Salt Lake City, UT, USA, June 2018.
- [48] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, Long Beach, CA, USA, June 2019.
- [49] C. Tang, W. Zhang, H. Li et al., "CNN-based qualitative detection of bone mineral density via diagnostic CT slices for osteoporosis screening," *Osteoporosis International*, vol. 32, no. 5, pp. 971–979, 2021.
- [50] Z. Wei, P. Yuhan, J. Jianhang et al., "RMSDSC-Net: a robust multiscale feature extraction with depthwise separable convolution network for optic disc and cup segmentation," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 11482–11505, 2022.
- [51] W. Wang, W. Zhou, J. Ji et al., "Deep sparse autoencoder integrated with three-stage framework for glaucoma diagnosis," *International Journal of Intelligent Systems*, vol. 37, no. 10, pp. 7944–7967, 2022.
- [52] K. Lei, A. B. Syed, X. Zhu, J. M. Pauly, and S. S. Vasanawala, "Artifact-andcontent-specific quality assessment for MRI with image rulers," *Medical Image Analysis*, vol. 77, Article ID 102344, 2022.
- [53] Y. Yi, C. Guo, Y. Hu, W. Zhou, and W. Wang, "BCR-UNet: Bi-directional ConvLSTM residual U-Net for retinal blood vessel segmentation," *Frontiers in Public Health*, vol. 10, pp. 1056226–1056313, 2022.
- [54] K. Lei, A. B. Syed, X. Zhu, J. M. Pauly, and S. V. Vasanawala, "Automated MRI field of view prescription from region of interest prediction by intra-stack attention neural network," *Bioengineering*, vol. 10, no. 1, p. 92, 2023.
- [55] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, "A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.