

Article

Prior Knowledge-Guided CNN-Swin Transformer Hybrid Network for Osteonecrosis of the Femoral Head in MRI

Cheng Yang  and Meiling Wang *

College of Engineering Science and Technology, Shanghai Ocean University, Shanghai 201306, China; m230851541@st.shou.edu.cn

* Correspondence: mlwang@shou.edu.cn

Abstract

Accurate JIC (Japanese Investigation Committee) classification of osteonecrosis of the femoral head (ONFH) is critical for collapse risk prediction and hip-preserving treatment. However, clinical classification faces challenges: indistinct lesion boundaries, limited annotated medical data, and the black-box inference issue of purely data-driven deep learning models. To address these, a Prior Knowledge-Guided CNN-Swin Transformer Hybrid Network (PGCT-Net) is proposed for high-accuracy classification with interpretable decision support. A cascaded dual-branch structure is adopted: the CNN branch extracts fine-grained local features from MRI images, while the Swin Transformer branch captures multi-scale global semantics and long-range dependencies between necrotic lesions and the acetabular weight-bearing region. A lesion mask-guided learning module injects expert-annotated clinical prior knowledge to focus the model on pathological regions and suppress background interference. Grad-CAM is used to visualize attention distribution for better interpretability. The network is trained end-to-end with a composite loss function combining cross-entropy loss and L1 sparse regularization. On the JLU-ONFH dataset, PGCT-Net achieves 94.38% accuracy, 94.15% F1-score and 93.97% AUC, significantly outperforming mainstream models. Cross-task validation on the BT dataset verifies the architecture's generalizability. This work provides an effective, interpretable scheme for ONFH JIC classification, with promising clinical auxiliary diagnosis potential.

Keywords: ONFH; MRI; deep learning; image classification; Swin Transformer; interpretability; prior knowledge

1. Introduction

In recent years, artificial intelligence technologies centered on deep learning have achieved breakthrough advancements in the field of medical image-assisted diagnosis. They have become core tools for improving the efficiency of disease diagnosis and reducing subjective diagnostic errors, and are widely applied in the full-range clinical scenarios of multimodal images, including CT, MRI and ultrasound, such as lesion detection, classification and grading, as well as prognosis evaluation [1]. In the field of diagnosis and treatment for musculoskeletal diseases, deep learning-based image analysis techniques can extract subtle pathological features that are difficult to recognize by the naked eye. They provide objective and quantitative auxiliary decision-making support for early disease diagnosis, precise classification and individualized treatment, and have emerged as a research hotspot in the interdisciplinary field of orthopedic clinical practice and medical imaging.



Received: 18 April 2026

Revised: 5 May 2026

Accepted: 6 May 2026

Published: 9 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Osteonecrosis of the Femoral Head (ONFH) is a progressive and disabling musculoskeletal disease triggered by the interruption or damage of the blood supply to the femoral head, which causes necrosis of osteocytes and bone marrow components, and further leads to structural degeneration of the femoral head [2]. The global incidence of this disease keeps rising, and it mainly occurs in young and middle-aged adults aged 20 to 50, with clinical features of high concealment, rapid progression and high disability rate [3,4]. Without accurate diagnosis and intervention in the early stage, more than 80% of patients will develop femoral head collapse within 1 to 4 years and eventually have to undergo total hip arthroplasty to restore joint function. For young and middle-aged patients, premature replacement will result in multiple difficult revision surgeries, imposing a heavy burden on patients' families and the social medical system. Therefore, realizing precise classification and diagnosis of ONFH in the golden window before femoral head collapse is of core clinical significance for predicting collapse risk, formulating individualized hip-preserving plans and improving the long-term prognosis of patients [5].

MRI can clearly depict key pathological signs, including bone marrow edema, the boundary of necrotic lesions and reactive sclerotic bands, which serve as the imaging gold standard for the early diagnosis and condition evaluation of ONFH [6,7]. In clinical practice, the evaluation of ONFH mainly includes two core dimensions: staging and classification. Staging systems represented by the ARCO system are primarily applied to assess the disease progression and femoral head collapse [8,9]. In contrast, the JIC classification system proposed by the Japanese Orthopaedic Association (JOA) focuses on the spatial location, the extent of the involvement of necrotic lesions, and their relative positional relationship with the acetabular weight-bearing area. The correlation between JIC classification outcomes, the risk of femoral head collapse, and the prognosis of hip-preserving treatment has been widely validated in multicenter clinical studies, making the JIC system a core basis for formulating clinical hip-preserving treatment strategies [10]. The JIC classification divides ONFH into four subtypes: A, B, C1 and C2. Among them, patients with type C2 have a femoral head collapse rate as high as 84.8%, representing the key population for early clinical intervention. There are significant differences in treatment regimens and prognoses among different subtypes, and accurate JIC classification is a prerequisite for successful hip-preserving treatment [11].

However, the current clinical JIC classification of ONFH still presents two major core challenges. First, classification diagnosis relies heavily on clinicians' experience, leading to significant intra-observer and inter-observer variability in judging lesion boundaries and the scope of the weight-bearing area. Multicenter studies have shown that the classification consistency among physicians from hospitals of different tiers is only 65–78%, which is prone to misjudgment [12]. Second, the imaging features of early ONFH lesions are subtle, and the differences between subtypes are only reflected in millimeter-level positional variations in lesion boundaries. The rate of missed diagnosis and misdiagnosis by junior physicians in primary hospitals exceeds 30% [13].

To address these issues, scholars worldwide have conducted extensive deep learning-based studies on ONFH imaging diagnosis. Convolutional Neural Networks (CNNs) serve as the mainstream technical solution, which can automatically extract fine-grained features such as lesion texture and boundaries, presenting satisfactory performance in ONFH lesion detection and staging classification [14]. However, the inherent limitation of the local receptive field in CNNs restricts them from effectively capturing the relative spatial relationship between lesions and the acetabular weight-bearing area. CNNs also lack the capability to extract fine-grained features of early tiny lesions, are prone to overfitting on small-sample medical datasets, and thus have constrained generalization ability [15]. Vision Transformer architectures represented by Swin Transformer can effectively model

the global spatial distribution and long-range dependencies of lesions via the global self-attention mechanism, exhibiting superior performance to traditional CNNs in identifying advanced ONFH lesions [16]. Nevertheless, they are weaker than CNNs in capturing local fine-grained features, with low sensitivity to early inconspicuous lesions. Furthermore, they tend to learn irrelevant background pseudo-features on small-sample datasets, leading to poor training stability and limited generalization and robustness [17].

To compensate for the inherent drawbacks of single architectures, existing studies have begun to explore the application of CNN-Transformer hybrid architectures in medical image classification. By integrating the strong local feature representation capability of CNNs and the global semantic modeling ability of Transformers, breakthroughs in diagnostic performance have been achieved in various medical image tasks [18,19], which also provides a new perspective for optimizing ONFH imaging diagnosis models. However, current hybrid architecture studies still suffer from critical limitations. Most adopt a pure data-driven training paradigm, lacking pathological prior knowledge and diagnostic logical constraints that align with clinical diagnostic criteria. This not only makes it difficult to conduct stable and efficient feature learning in small-sample scenarios but also results in insufficient interpretability and reliability of the model, rendering it unable to directly provide credible auxiliary support for clinical hip-preserving decision-making [20].

To address the aforementioned clinical challenges and limitations of existing studies, a Prior Knowledge-Guided CNN-Swin Transformer Hybrid Network (PGCT-Net) is proposed for the accurate JIC classification of osteonecrosis of the femoral head (ONFH) using MRI images. First, a cascaded fusion architecture integrating CNN and Swin Transformer is constructed. Fine-grained local features such as lesion edges and textures are extracted by the CNN module, while global semantic features including necrotic lesion scope are modeled by the Swin Transformer module, achieving deep complementary fusion of local and global features. Second, a clinical lesion mask-guided learning mechanism is introduced, in which expert-annotated clinical prior knowledge of lesions is deeply embedded into the full pipeline of model feature learning. This mechanism enables the model to focus on the core regions of necrotic lesions and suppress invalid feature learning from irrelevant background areas, thereby enhancing the feature representation capability and classification performance of the model. Finally, attention distribution on lesion regions is visualized via Grad-CAM mapping, which significantly improves the interpretability and reliability of model decision-making. Extensive experiments, including ablation studies, comparative experiments, feature visualization and interpretability analysis, are conducted on the publicly available JLU-ONFH MRI dataset and a public brain tumor dataset. The classification performance, cross-dataset robustness on medical images and clinical interpretability of PGCT-Net are systematically validated. This work aims to provide a reliable and interpretable artificial intelligence-aided diagnostic tool for the accurate classification and individualized hip-preserving treatment of ONFH.

The core contributions of this study are summarized as follows:

- (1) A PGCT-Net hybrid architecture with cascaded fusion of CNN and Swin Transformer is proposed, which realizes the deep complementary fusion of local and global features. It solves the core problem that single architectures can hardly simultaneously achieve the recognition of early subtle lesions and the modeling of spatial relationships among lesion regions and significantly improves the diagnostic accuracy of ONFH JIC classification.
- (2) A clinical lesion mask-guided learning mechanism is designed to embed expert-annotated clinical prior knowledge into the entire feature learning process. The model is guided to focus on the core diagnostic regions of necrotic lesions and suppress redundant feature extraction from background regions, leading to improved classification accuracy.

(3) Grad-CAM visualization is utilized to conduct heatmap-based feature attribution analysis for the model, which enhances the transparency of the model's decision-making process. Comprehensive performance validation is also carried out on multiple datasets. This method can intuitively reveal the key lesion regions attended by the model, effectively improving the practical applicability of the proposed approach in clinical scenarios.

2. Related Work

2.1. Research on Intelligent Medical Image Diagnosis of ONFH

Deep learning-based imaging analysis technology has emerged as the mainstream research direction in the intelligent diagnosis of ONFH. Most existing studies focus on lesion detection and staging diagnosis of ONFH. For instance, relevant studies constructed ONFH lesion segmentation models based on Convolutional Neural Networks (CNNs) to achieve the automatic extraction of necrotic regions [21]; other studies realized automatic classification of ONFH ARCO staging based on the ResNet architecture, yielding promising diagnostic performance [22].

Existing studies on the typed diagnosis of ONFH remain relatively scarce and face obvious performance bottlenecks. In clinical practice, the JIC classification is directly correlated with the risk of femoral head collapse and hip-preserving treatment plans. Its core diagnostic focus lies in determining the relative positional relationship between the lesion and the weight-bearing area of the acetabulum, which imposes extremely high requirements on the model's ability of global spatial modeling and fine-grained feature extraction. Existing CNN-based methods are limited by local receptive fields, making it difficult to effectively model the long-range spatial dependencies between the lesion and the weight-bearing area, and thus insufficiently distinguishing between subtypes such as C1/C2 that only have millimeter-level positional differences. Although Transformer-based methods have the advantage of global modeling, they are prone to feature drift on small-sample ONFH datasets, resulting in low sensitivity in identifying early occult lesions. In addition, most existing deep learning models adopt a pure data-driven mode, which suffers from the obvious problem of black-box decision-making. The decision-making process lacks interpretability, making it difficult to ensure the reliability and credibility required for clinical application [23].

2.2. Medical Image Classification Methods Based on CNN and Transformer

Convolutional Neural Networks (CNN) serve as a classic architecture in the field of medical image classification. By virtue of local weight sharing and sliding window mechanisms, they present prominent advantages in fine-grained local feature extraction. Since AlexNet [24] was proposed, classic CNN architectures including VGG [25], ResNet [26] and MobileNet [27] have been widely employed in medical image classification tasks. To tackle the overfitting problem in small-sample scenarios of medical images, Ferdous et al. [28] combined depthwise separable convolution with the CBAM attention mechanism to efficiently extract and refine spatial and channel features, which effectively mitigated overfitting on small-sample datasets. Li et al. [29] adopted a parameterized Gaussian distribution-driven data cleaning strategy to improve data quality and optimized non-pretrained architectures to adapt to resource-constrained environments, elevating the classification accuracy of the model. Nevertheless, the inherent inductive bias and local receptive field limitation of CNNs make it hard for them to effectively model long-range dependencies in images and fail to precisely capture the relative spatial positional relationship between ONFH lesions and the acetabular weight-bearing area. This is the core cause of their restricted performance in JIC classification tasks.

The emergence of Vision Transformer (ViT) [30] provides a new technical path for global feature modeling of medical images. Through the self-attention mechanism, it can effectively capture long-range dependencies at arbitrary positions in images and is significantly superior to traditional CNNs in global semantic modeling. Among them, Swin Transformer [31] further optimizes the computational efficiency and multi-scale feature extraction capability of ViT via a hierarchical window attention mechanism and a multi-scale feature pyramid, and has been widely used in medical image classification, lesion detection and other tasks. Chen et al. [32] proposed a benchmark-driven clinical decision framework for multi-class diagnosis of middle ear diseases, verifying the superiority of Swin Transformer in diagnostic accuracy and stability; Xu et al. [33] proposed an intelligent fracture classification method based on Swin Transformer, realizing high-precision automatic fracture typing and radiological evaluation. Nevertheless, the Transformer architecture is weaker than CNNs in capturing local fine-grained texture features, with insufficient recognition sensitivity for early occult ONFH lesions. Meanwhile, the unconstrained self-attention mechanism easily leads the model to learn background pseudo-features irrelevant to diagnosis, resulting in poor training stability on small-sample ONFH datasets with scarce annotated samples.

2.3. Application of CNN-Transformer Hybrid Architecture in Medical Imaging

To compensate for the inherent drawbacks of standalone CNN and Transformer architectures, the CNN-Transformer hybrid architecture has emerged as a research hotspot in medical imaging. By integrating the fine-grained local feature representation capability of CNNs and the global semantic modeling capability of Transformers, this architecture realizes deep complementary fusion of local and global features and has achieved performance breakthroughs in a variety of medical image tasks.

Wang et al. [34] proposed MFF-ClassificationNet, a dual-branch parallel hybrid architecture. It employs the CNN branch to capture local detailed texture features and the Transformer branch to model global dependencies, which significantly strengthens the model's ability to extract multi-scale features and contextual information, thus improving diagnostic accuracy. Qezelbash-Chamak et al. [35] presented a learnable hybrid fusion architecture that optimizes image classification performance by adaptively integrating the strengths of the two architectures. However, most existing research on hybrid architectures follows a purely data-driven training paradigm, without pathological prior knowledge and diagnostic logical constraints that conform to clinical diagnostic standards during training. This not only hinders stable and efficient feature learning in small-sample scenarios but also leads to an uninterpretable decision-making process of the model. Consequently, it is difficult to gain the trust of clinicians and cannot be directly applied to clinical auxiliary decision-making.

3. Materials and Methods

3.1. PGCT-Net

Aiming at the challenges in the imaging classification task of early-stage osteonecrosis of the ONFH, including tiny lesion regions, scarce clinical samples and insufficient interpretability of traditional models, this paper proposes a Prior Knowledge-Guided CNN-Swin Transformer Hybrid Network (PGCT-Net). The overall architecture of the model is illustrated in Figure 1.

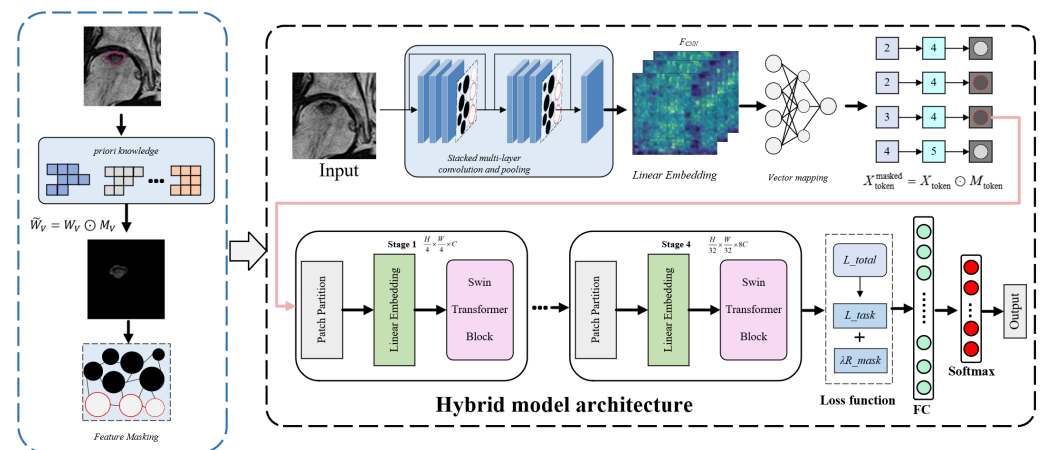


Figure 1. The structure illustration of PGCT-Net.

Guided by the spatial prior constraints introduced by lesion mask learning as the core, the model deeply embeds such constraints into the whole process of shallow local feature extraction via CNN and deep global feature modeling via Swin Transformer. The spatial prior constraints accurately guide CNN to concentrate on lesion regions and suppress the interference from background noise, while empowering it to efficiently capture fine-grained imaging representations including the edges and textures of tiny lesions. On this basis, relying on the local features optimized by mask constraints, Swin Transformer further mines the global contextual correlations and multi-scale semantic information of lesions, realizing the deep coupling between local fine-grained features and global semantic features.

3.2. Prior Knowledge Embedding

Mask learning is a general technique that embeds trainable mask parameters into deep learning models, aiming to achieve dynamic and adaptive regulation of the model's internal components. Its core idea is to automatically identify and strengthen task-critical information flows through a learnable gating mechanism, while suppressing redundant or noisy components, so as to improve the model's performance, efficiency and interpretability.

In Transformer-based mask learning models, the mask is usually designed as a matrix with the same shape as the corresponding trainable parameter. For the value weight matrix W_V in the self-attention module, a learnable mask M_V is introduced. The mask acts on the original parameter via element-wise multiplication to generate the effective parameter used in actual forward propagation:

$$\tilde{W}_V = W_V \odot M_V, \quad (1)$$

where $\odot$ denotes the element-wise multiplication. At the initial stage of training, the mask is generally initialized as an all-one matrix to ensure the normal startup of the model. As training iterations proceed, mask parameters and the backbone weights of the model are jointly optimized end-to-end via gradient descent. Meanwhile, sparsity regularization constraints drive the mask values corresponding to unannotated regions to gradually approach zero, which softly masks background interference and achieves automatic and precise screening of features associated with annotated regions.

Taking ONFH as an example, for the input raw hip joint MRI images, we constructed a standardized manual annotation workflow to generate the binary lesion mask strictly pixel-aligned with the original images, which serves as the core clinical prior knowledge for model training. The specific annotation protocol is as follows:

(1) Annotator configuration: The annotation work was completed by two attending radiologists with 8 years and 10 years of clinical experience in musculoskeletal MRI diagnosis respectively, and a chief orthopedic surgeon with 15 years of clinical experience in ONFH diagnosis and treatment provided professional guidance on lesion boundary definition.

(2) Annotation standard: The annotation scope strictly follows the JIC classification diagnostic criteria, including the necrotic core area, reactive sclerotic band, cystic degeneration area, and bone marrow edema area related to ONFH lesion classification.

(3) Quality control and consistency assessment: A double-blind independent annotation strategy was adopted. The two radiologists completed the annotation of all images independently, and the inter-rater agreement was evaluated using Cohen's kappa coefficient. The overall kappa coefficient of the annotation results was 0.917, indicating excellent inter-rater consistency. For the lesion regions with annotation discrepancies, the chief orthopedic surgeon and the two radiologists conducted a joint review to reach a final consensus annotation result.

(4) Mask post-processing: The consensus annotation results were generated into binary mask images, where the lesion region was assigned a value of 1 and the non-lesion background region was assigned a value of 0. Bilinear interpolation and patch-level pooling were used to achieve spatial alignment between the mask and multi-scale features of the model. The generation process is illustrated in Figure 2.

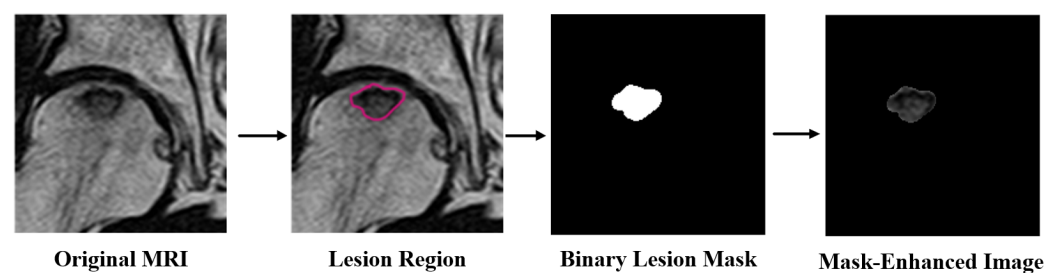


Figure 2. Flowchart of mask generation.

Guided by clinical prior knowledge, the mask divides the image into lesion foreground and irrelevant background. It serves as spatial prior knowledge that can be directly embedded into model training and is only used during the training phase. In the mask matrix, physician-annotated regions such as the necrotic core area, sclerotic zone, cystic degeneration, and bone marrow edema are uniformly assigned a value of 1, representing effective feature regions highly correlated with staging diagnosis. Non-target regions including the acetabular margin, joint effusion, muscles, and fat are assigned a value of 0, representing redundant background regions that are prone to interference.

To achieve spatial alignment between the mask and multi-scale features, the feature-level mask is resized via bilinear interpolation for scale adaptation, and the token-level mask is generated by average pooling within each patch, followed by binarization, ensuring strict spatial consistency.

Such spatial constraints filter out redundant information, forcing the model to learn key discriminative features only in the core lesion areas, thereby significantly improving the specificity and accuracy of feature learning. During training, the prior mask acts as supervised guidance to help the model solidify a lesion-focused attention mechanism. After training is completed, the model can automatically identify and focus on lesion regions directly from raw MRI images. No mask input, manual annotation, or additional segmentation is required in clinical practice, which avoids extra clinical workload and error propagation caused by automatic segmentation, ensuring high clinical feasibility and diagnostic efficiency.

3.3. CNN Shallow Local Feature Extraction Module

Convolutional Neural Networks (CNNs) demonstrate remarkable applicability in medical image analysis, owing to their strong ability to capture local image features and the advantages of hierarchical feature learning. Its model architecture is shown in Figure 3.

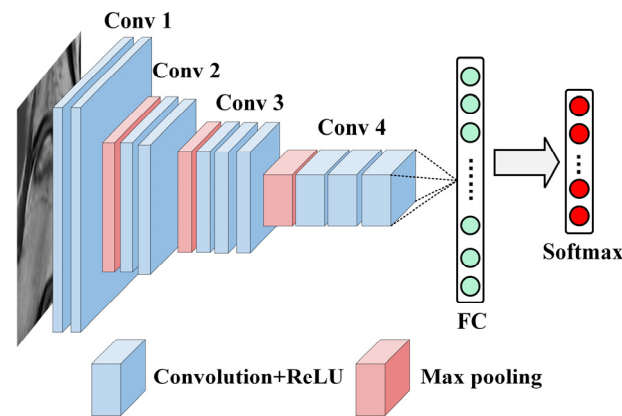


Figure 3. The structure illustration of CNN.

Taking the mask-enhanced image as input, fine-grained extraction of shallow local features in osteonecrosis of the femoral head regions is realized via multi-layer convolution and pooling operations. Meanwhile, the whole process is guided by the spatial constraints of the prior mask, ensuring that feature learning always concentrates on key lesion regions.

This module consists of 4 alternately stacked convolution–activation–pooling units. Let X_{mask} denote the mask-preprocessed input image, then its forward propagation process at the l -th convolutional layer can be expressed as:

$$F_l = \sigma(K_l * F_{l-1} + b_l), \quad (2)$$

where $*$ denotes the convolution operation, K_l and b_l represent the convolution kernel weights and bias term of the l -th layer respectively, $\sigma(\cdot)$ is the ReLU nonlinear activation function, and F_{l-1} denotes the output feature of the previous layer.

To be consistent with the prior knowledge embedding mechanism, a feature-level mask constraint is introduced after the output of each convolutional layer:

$$F_l^{\text{masked}} = F_l \odot M_l, \quad (3)$$

where M_l denotes the feature-level mask after size adaptation, and $\odot$ represents element-wise multiplication. This operation can enhance the feature response of lesion regions, suppress redundant background information, and enable convolution kernels to focus more on learning fine-grained representations such as textures, trabecular bone structures, boundary contours and grayscale gradient variations in lesion areas.

The feature-level mask $M_{feat}^{(l)}$ at layer l is obtained by downsampling the physician mask M_{phy} via bilinear interpolation to match the spatial size of F_l :

$$M_{feat}^{(l)} = \text{Resize}\left(M_{phy}, (H_l, W_l)\right), \quad (4)$$

where H_l , W_l are the height and width of the l -th convolutional feature map. This ensures strict spatial alignment between the mask and feature responses.

The spatial size of the output feature map from the convolutional layer can be calculated by the following equation:

$$W_{\text{out}} = \frac{W_{\text{in}} + 2P - K}{S} + 1, \quad (5)$$

where W_{in} denotes the size of the input feature map, P represents the zero-padding width, K is the convolution kernel size, and S stands for the sliding stride.

Each convolutional unit is followed by a max-pooling layer for downsampling to further compress feature dimensions and enhance feature robustness. The max-pooling operation is defined as:

$$P_l(i, j) = \max_{(m, n) \in R_{i, j}} F_l^{\text{masked}}(m, n), \quad (6)$$

where $R_{i, j}$ denotes the receptive field region corresponding to the pooling window. The pooling operation retains salient lesion features while reducing the number of parameters and expanding the receptive field of subsequent convolutional layers.

After feature extraction, the module outputs a deep local feature map F_{CNN} with a size of $H \times W \times C$. Specifically, the CNN module adopts 3×3 convolution kernels with same padding, and the number of channels is set to 32, 64, 128, and 256 sequentially, resulting in a final local feature map of $14 \times 14 \times 512$ dimensions. This feature map not only retains the fine structural information of early osteonecrosis of the femoral head lesions but also suppresses redundant features and preserves critical features under mask constraints, providing high-quality and highly targeted feature input for the global semantic modeling of the subsequent Swin Transformer.

3.4. Swin Transformer Deep Multi-Scale Feature Encoding Module

Swin Transformer is a hierarchical Vision Transformer architecture designed to address the problems of excessively high computational complexity and the lack of multi-scale feature representation in traditional Vision Transformers (e.g., ViT) when processing high-resolution images. Its model structure is shown in Figure 4.

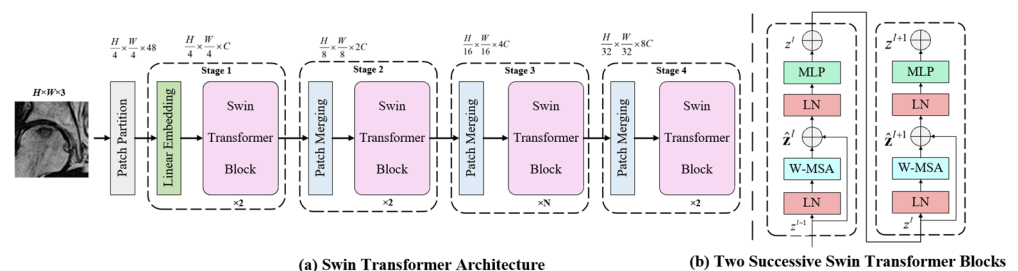


Figure 4. The structure illustration of Swin Transformer.

Taking over the feature map F_{CNN} extracted by the shallow CNN and refined under mask constraints, the module performs in-depth modeling of multi-scale global semantic features in osteonecrosis of the femoral head regions based on the hierarchical window attention mechanism. Meanwhile, it continues the prior knowledge embedding strategy to ensure that global feature learning always concentrates on the core lesion areas, compensating for the limitation of CNN that only captures local features.

First, F_{CNN} is processed through Patch Partition and Linear Embedding, and mapped into the initial token sequence X_{token} compatible with Transformer. The mapping process can be expressed as:

$$X_{\text{token}} = \text{Linear}(\text{PatchPartition}(F_{\text{CNN}})). \quad (7)$$

To continue the prior knowledge constraint, a mask filtering operation is introduced after token mapping to retain only the tokens corresponding to lesion regions and suppress the feature response of invalid tokens in background regions. The filtering process is given by:

$$X_{\text{token}}^{\text{masked}} = X_{\text{token}} \odot M_{\text{token}}, \quad (8)$$

where M_{token} denotes the token-level mask generated from the feature-level mask M_{feat} via patch partitioning, and $\odot$ represents element-wise multiplication, ensuring that the subsequent attention computation only focuses on lesion-related tokens.

The token-level mask M_{token} is derived from the physician mask by patch-wise average pooling followed by binarization:

$$M_{\text{token}}(i, j) = \mathbf{1}(\frac{1}{p^2} \sum_{u=1}^p \sum_{v=1}^p M_{\text{phy}}(i \cdot + u, j \cdot + v) > \theta), \quad (9)$$

where p is the patch size, $\theta = 0.5$ is the binarization threshold. The attention mask M_{attn} is defined as an additive bias matrix that disables cross-token attention between non-lesion tokens:

Subsequently, $X_{\text{token}}^{\text{masked}}$ is input into 4 stages for hierarchical encoding, where each stage consists of a Patch Merging layer and stacked Swin Transformer Blocks.

The core of the Swin Transformer Block lies in an alternating structure of Window Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA). Combined with the constraint of the attention mask M_{attn} , it only establishes long-range dependencies within the lesion-related token regions. The calculation formulas are given as follows:

$$W - \text{MSA}(X) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \odot M_{\text{attn}}, \quad (10)$$

SW-MSA enables cross-window lesion feature interaction via window shifting, which compensates for the limitation of W-MSA. The Patch Merging layer achieves downsampling through concatenation of adjacent tokens and linear projection, constructing a multi-scale feature pyramid. After encoding through four stages, the module outputs a deep multi-scale feature map F_{Swin} that fuses local and global features. The output feature sizes of each stage are as follows:

$$H_t = \frac{H}{2^t}, W_t = \frac{W}{2^t}, C_t = C \times 2^t \quad (t = 1, 2, 3, 4). \quad (11)$$

Specifically, the Swin Transformer is configured with a window size of 7, an embedding dimension of 96, 3, 6, 12, 24 attention heads in the four stages, and stacked block numbers of [2, 2, 6, 2]. Such features encompass both the fine-grained structural details and global semantic information of lesions, providing highly discriminative support for subsequent classification decisions and achieving the deep integration of prior knowledge and global feature modeling.

3.5. End-to-End Joint Optimization Strategy

The model adopts an end-to-end training paradigm, which integrates the mask prior constraint, the feature extraction weights of CNN and Swin Transformer, as well as the parameters of the classification head into a unified optimization framework. Based on the joint loss function for mask learning, the total loss function is formulated as:

$$L_{\text{total}} = L_{\text{task}} + \lambda R_{\text{mask}}, \quad (12)$$

where L_{task} denotes the cross-entropy loss for multi-stage classification of osteonecrosis of the femoral head, R_{mask} represents the L1 sparsity regularization term of the mask matrix. λ is the regularization hyperparameter, which is set to 0.005 in this study. Effective sparsity constraints on the mask are achieved while high classification accuracy is maintained.

The training inputs consist of the raw MRI image X and the physician mask M_{phy} . The forward pass process during the training phase is as follows:

- (1) Resize M_{phy} to multi-scale feature-level masks $M_{feat}^{(l)}$ and token-level mask $M_{token}^{(l)}$.
- (2) Compute masked CNN features: $F_l^{masked} = \text{Conv}(F_{l-1}) \odot M_{feat}^{(l)}$.
- (3) Compute masked Swin tokens: $X_{token}^{masked} = X_{token} \odot M_{token}^{(l)}$.
- (4) Compute W-MSA using the additive attention mask M_{attn} .
- (5) Final classification loss combined with an L1 sparsity regularization term applied to the learnable mask M_{learn} .

All parameters of the model are jointly optimized through backpropagation and gradient descent algorithms. While minimizing classification loss, the mask module is continuously driven to strengthen feature activation in lesion regions and suppress invalid information in background regions. This makes the feature learning logic of the model highly consistent with the diagnostic logic of clinicians, further improving the classification accuracy of the model.

4. Dataset Description

In this study, two public medical MRI datasets were employed to evaluate the model performance, namely the Jilin University Osteonecrosis of the Femoral Head dataset (JLU-ONFH) and the brain tumor (BT) dataset. All MRI examinations included in this study were performed on 3.0 T clinical scanners with standardized imaging protocols, and the detailed sequence parameters are described separately for each dataset below. All images underwent conventional data preprocessing (including size normalization and grayscale standardization) and were uniformly resized to an input resolution of 224×224 pixels. Representative examples of the two datasets are illustrated in Figure 5.

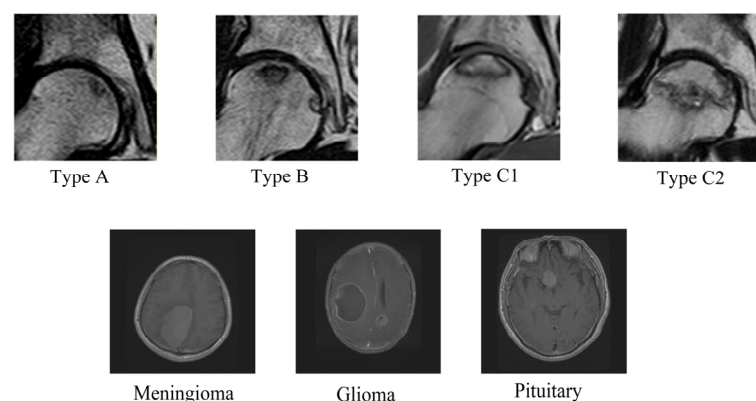


Figure 5. Examples of JLU-ONFH and BT datasets.

4.1. JLU-ONFH Dataset

The JLU-ONFH dataset (public download link: <https://github.com/sole12343/JLU-ONFH-Dataset> (accessed on 13 April 2026)) is a public ONFH MRI dataset constructed by the Department of Orthopedics and Radiology. The Second Bethune Hospital of Jilin University, and has been widely used in peer-reviewed ONFH intelligent diagnosis studies [36]. This study adopts the classification subset of the official dataset, which is specially organized for JIC classification tasks of ONFH. All images are T1-weighted spin-echo MRI images converted from original DICOM files, with all patient privacy information

completely desensitized and stored in PNG format. All MRI scans in this dataset were performed on the same 3.0 T Siemens Magnetom Verio scanner with a unified standardized imaging protocol to ensure full consistency of scanning settings across all examinations: repetition time (TR) of 400–600 ms, echo time (TE) of 10–20 ms, slice thickness of 4 mm, intersection gap of 1 mm, field of view (FOV) of 380×380 mm, and matrix size of 256×256 .

The final dataset used in this study contains 1448 mid-slice MRI images from 327 independent pathologically confirmed ONFH patients, with each patient contributing 2–6 valid mid-slice images of the affected hip joint. No patient overlap exists between different JIC classification subtypes. All images were independently annotated and classified into four ONFH subtypes according to the JIC classification criteria by two senior orthopedic radiologists, with the following distribution: 178 images of Type A, 221 images of Type B, 348 images of Type C1, and 701 images of Type C2. A mild class imbalance is observed in the original dataset, with Type C2 (highest collapse risk) as the majority class.

To completely eliminate data leakage and ensure the objectivity and reproducibility of model performance evaluation, the full data processing workflow strictly follows the gold-standard normative principle of patient-level independent splitting first, unified preprocessing second, and data augmentation exclusively for the training set, with the following standardized sequential steps:

Step 1: Patient-level independent dataset splitting with stratified sampling

First, the 327 independent patients were randomly divided into the training set (229 patients, 70%), validation set (65 patients, 20%), and test set (33 patients, 10%) at a ratio of 7:2:1, with stratified sampling based on JIC classification subtypes to maintain the original class distribution in each subset. We strictly enforced the patient-level non-overlap rule: all MRI slices from the same patient were exclusively assigned to only one of the three sets, with no cross-set patient overlap, slice duplication or information leakage between the training, validation and test sets.

After splitting, the original unmodified image distribution of each subset is as follows: training set—1014 original images (Type A: 125, Type B: 155, Type C1: 243, Type C2: 491), validation set—290 original images (Type A: 35, Type B: 44, Type C1: 70, Type C2: 141), test set—144 original images (Type A: 18, Type B: 22, Type C1: 35, Type C2: 69).

To ensure stable and unbiased performance evaluation on imbalanced data, we further performed class-balanced random sampling on the validation and test sets and finally obtained 100 samples per class for the validation set (400 total images) and 50 samples per class for the test set (200 total images). All samples in the validation and test sets retained the original unmodified MRI images, with no augmentation or artificial modification. The detailed distribution is summarized in Table 1.

Table 1. JLU-ONFH dataset description.

ONFH Subtype	Training Set	Validation Set	Test Set
Type A	450	100	50
Type B	450	100	50
Type C1	450	100	50
Type C2	450	100	50

Step 2: Unified data preprocessing for all subsets

All images in the training, validation and test sets underwent the same standardized preprocessing operations, including size normalization to 224×224 pixels and grayscale standardization. No differential preprocessing was performed for different sets to ensure complete consistency of image input specifications.

Step 3: Data augmentation exclusively for the training set after splitting

To alleviate the class imbalance issue and improve the generalization ability of the model, lightweight unsupervised data augmentation strategies, including random horizontal flipping, random rotation within $\pm 15^\circ$, and random brightness adjustment, were only applied to the original images in the training set after patient-level splitting. All augmented variants were generated only from the original images of patients in the training set, and augmented samples from the same original image, hip or patient were strictly confined to the training set, with no augmented samples entering the validation or test sets. After class-balanced augmentation, the number of training samples for each category was uniformly adjusted to 450, with a total of 1800 training samples.

To further rule out potential data leakage, we performed a strict two-step duplicate and near-duplicate check across the training, validation, and test sets. We performed pixel-wise comparison of all images, and confirmed that there were no identical images across different sets. We set a strict threshold of SSIM > 0.95 to identify near-duplicate images with high structural similarity, and the results showed that there were no near-duplicate images across the three sets.

The lesion mask annotation for prior knowledge-guided learning was completed based on the original MRI images of this dataset, following a standardized clinical annotation protocol as detailed in Section 3.2.

4.2. BT Dataset

The BT dataset (publicly available at https://figshare.com/articles/dataset/brain_tumor_dataset/1512427 (accessed on 13 April 2026)) is a widely used public brain tumor MRI dataset constructed by Cheng Jun et al. [37], which contains 3064 T1-weighted contrast-enhanced MRI images from multi-center clinical institutions (General Hospital of Tianjin Medical University and Nanfang Hospital). All images were acquired with a standardized 3.0 T MRI scanning protocol, with detailed image acquisition parameters and format conversion code provided in the official dataset documentation.

To ensure data quality and experimental reliability, we performed a strict multi-step data screening process on the original 3064 images, with the following exclusion criteria: (1) images with signal-to-noise ratio (SNR) below 20 dB; (2) images with severe motion artifacts or geometric distortion; (3) images with blurred lesion boundaries or incomplete anatomical structures; (4) images with inconsistent slice thickness or non-compliant scanning parameters; (5) duplicate images from the same patient. After screening, a total of 1800 high-quality images from 386 independent patients with consistent scanning settings were retained for subsequent experiments. These images are divided into three categories according to pathological tumor types: Meningioma (600 images from 127 patients), Glioma (600 images from 132 patients), and Pituitary tumor (600 images from 127 patients), with a perfectly balanced class distribution.

The standardized MRI scanning protocol of the retained images is as follows: repetition time (TR) of 400–700 ms, echo time (TE) of 10–25 ms, slice thickness of 5 mm, no intersection gap, field of view (FOV) of 220×220 mm, and matrix size of 256×256 .

The data processing workflow of the BT dataset completely follows the same gold-standard normative principles as the JLU-ONFH dataset to eliminate data leakage, with the following sequential steps:

Step 1: Patient-level independent dataset splitting with stratified sampling

First, the 386 independent patients were randomly divided into the training set (270 patients, 70%), validation set (77 patients, 20%), and test set (39 patients, 10%) at a ratio of 7:2:1, with stratified sampling based on tumor types to maintain balanced class distribution in each subset. We strictly enforced the patient-level non-overlap rule: all MRI slices from the same patient were exclusively assigned to only one of the three sets, with

no cross-set patient overlap, slice duplication or information leakage between the training, validation and test sets.

After splitting, the original unmodified image distribution of each subset is as follows: training set—1260 original images (Meningioma: 420, Glioma: 420, Pituitary: 420), validation set—360 original images (Meningioma: 120, Glioma: 120, Pituitary: 120), test set—180 original images (Meningioma: 60, Glioma: 60, Pituitary: 60).

To maintain consistency with the JLU-ONFH dataset experimental settings, we performed class-balanced random sampling on the validation and test sets (without any data augmentation), and finally obtained 100 samples per class for the validation set (300 total images) and 50 samples per class for the test set (150 total images). All samples in the validation and test sets retained the original unmodified MRI images, with no augmentation or artificial modification. The detailed distribution is summarized in Table 2.

Table 2. BT dataset description.

Label	Tumor Type	Training Set	Validation Set	Test Set
0	Meningioma	450	100	50
1	Glioma	450	100	50
2	Pituitary	450	100	50

Step 2: Unified data preprocessing for all subsets

All images in the training, validation and test sets underwent the same standardized preprocessing operations as the JLU-ONFH dataset, including size normalization to 224×224 pixels and grayscale standardization, to ensure complete consistency of input specifications across two datasets.

Step 3: Data augmentation exclusively for the training set after splitting

Lightweight data augmentation strategies completely consistent with the JLU-ONFH dataset were only applied to the original images in the training set after patient-level splitting. All augmented variants were generated only from the original images of patients in the training set, and strictly confined to the training set, with no augmented samples entering the validation or test sets. After augmentation, the number of training samples for each category remained 450, consistent with the JLU-ONFH dataset, to ensure fair cross-dataset comparison.

We performed the same two-step duplicate and near-duplicate check as the JLU-ONFH dataset and confirmed that there were no exact duplicate or near-duplicate ($\text{SSIM} > 0.95$) images across the training, validation, and test sets.

The lesion mask annotation for prior knowledge-guided learning was completed based on the original MRI images of this dataset, following a standardized clinical annotation protocol as detailed in Section 3.2.

5. Experiments

5.1. Evaluation Metrics

In the field of medical image classification, the problem of class imbalance is prominent, and categories with a small number of samples often correspond to subtypes with high clinical risk and critical intervention value. Therefore, a single evaluation metric cannot accurately measure the comprehensive performance of the model. For this 4-class ONFH JIC classification multiclass task and 3-class brain tumor classification task, we select accuracy (ACC), macro-averaged F1-score, and macro-averaged area under the ROC curve (AUC) as the core evaluation metrics, which can comprehensively evaluate the model's classification performance on both majority and minority classes in imbalanced multiclass scenarios. All samples in this study are clinical MRI images of patients with confirmed diseases (ONFH

or brain tumor), and no healthy control samples are included in the classification task. All metrics are calculated based on the classification of disease subtypes, without involving the discrimination between disease and healthy controls. The specific definitions are given as follows.

$$ACC = \frac{\sum_{i=1}^C TP_i}{N}, \quad (13)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}, \quad (14)$$

$$Macro - Recall = \frac{1}{C} \sum_{i=1}^C Recall_i, \quad (15)$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}, \quad (16)$$

$$Macro - Precision = \frac{1}{C} \sum_{i=1}^C Precision_i, \quad (17)$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}, \quad (18)$$

$$Macro - F1 = \frac{1}{C} \sum_{i=1}^C F1_i, \quad (19)$$

where N denotes the total number of samples in the test set, and C denotes the total number of classification categories ($C = 4$ for the ONFH task, $C = 3$ for the brain tumor task). TP_i (True Positive): The number of samples that truly belong to category i and are correctly classified as category i , TN_i (True Negative): The number of samples that do not belong to category i and are correctly classified as non- i categories, FP_i (False Positive): The number of samples that do not belong to category i but are misclassified as category i , FN_i (False Negative): The number of samples that truly belong to category i but are misclassified as non- i categories. For the AUC, we adopt the macro-averaged one-vs-rest (OVR) AUC, which is calculated by fitting the ROC curve for each subtype against all other subtypes separately, then averaging the AUC values of all subtypes. This strategy is widely recommended for imbalanced multi-class medical image classification tasks, as it treats all subtypes equally and reflects the overall discriminative ability of the model.

5.2. Experimental Setup

To ensure effective and stable model training, we adopt consistent and unified experimental settings for fair evaluation of medical image classification performance. All experiments are implemented on the Windows 11 operating system with an NVIDIA RTX 4060 (8 G) GPU for acceleration. The framework is built on PyTorch 1.9.0+cu111 and coded with Python 3.9.19 (64-bit), supported by CUDA 11.1 for parallel computing. The key hyperparameters are configured as follows: the batch size is set to 16; the model is optimized by the Adam optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-5} , and the learning rate is adjusted by the cosine annealing scheduler during training. The network parameters are initialized with the Xavier method, and the model is trained for 100 epochs. We adopt an early stopping strategy by saving the optimal model based on the validation AUC to avoid overfitting.

5.3. Experimental Results and Discussion

5.3.1. Experimental Results on the Dataset JLU-ONFH

To verify that the proposed method can effectively distinguish MRI images of osteonecrosis of the ONFH, the constructed JLU-ONFH dataset is fed into the PGCT-Net model for training. The number of training epochs is set to 100. The experimental results are shown in Figures 6 and 7, and the detailed per-class diagnostic performance metrics are summarized in Table 3.

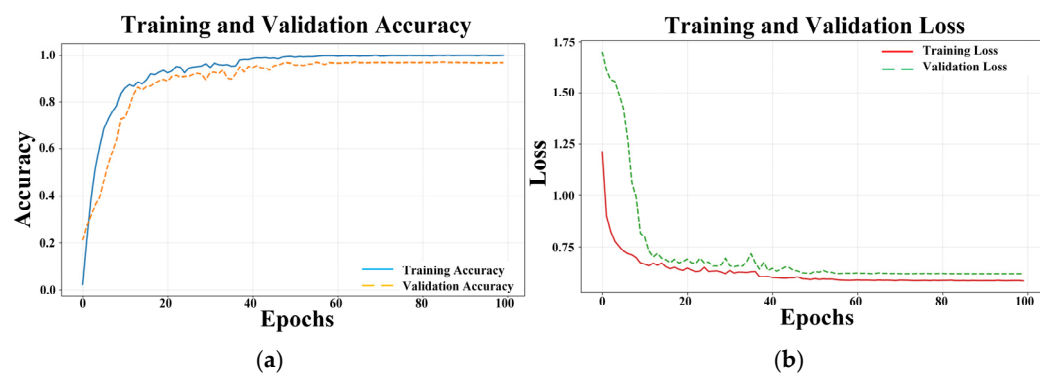


Figure 6. Convergence curves of the PGCT-Net on the dataset JLU-ONFH. (a) Accuracy curve; (b) loss curve.

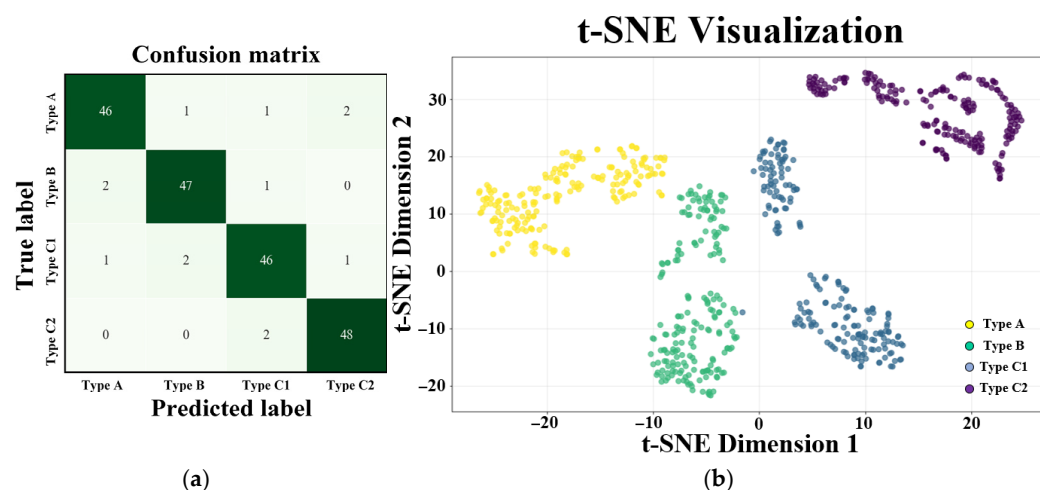


Figure 7. Result visualization of the PGCT-Net on the dataset JLU-ONFH. (a) Confusion matrix; (b) t-SNE visualization result.

Table 3. Per-class diagnostic performance metrics of PGCT-Net on the JLU-ONFH test set (mean $\pm$ 95% CI).

ONFH JIC Subtype	Recall (%)	Precision (%)	F1 (%)	AUC (%)
Type A	96.13 $\pm$ 2.15	97.94 $\pm$ 1.54	96.95 $\pm$ 1.23	99.19 $\pm$ 0.58
Type B	94.41 $\pm$ 2.58	96.89 $\pm$ 1.76	95.41 $\pm$ 1.45	98.74 $\pm$ 0.72
Type C1	92.06 $\pm$ 2.94	95.81 $\pm$ 1.92	93.86 $\pm$ 1.67	97.87 $\pm$ 0.85
Type C2	96.15 $\pm$ 2.15	94.10 $\pm$ 2.03	95.03 $\pm$ 1.34	90.04 $\pm$ 1.12
Macro-averaged	94.69 $\pm$ 1.85	96.19 $\pm$ 1.31	94.14 $\pm$ 0.67	93.96 $\pm$ 0.91

Figure 6a shows the diagnostic accuracy curves of the model. In the first 20 training epochs, the model exhibits efficient convergence characteristics, and the training accuracy and validation accuracy increase synchronously and significantly. After entering the stable stage, the training accuracy gradually approaches 100%, while the validation accuracy remains stable at approximately 94.38%, indicating that the model has excellent fitting

ability. Figure 6b presents the corresponding loss function curves. Both the training loss and validation loss decrease rapidly within the first 20 epochs. Subsequently, the training loss gradually converges to a level close to 0, and the validation loss also stabilizes in a low numerical range without obvious fluctuations after convergence, which further corroborates the stability and reliability of the model training process.

Figure 7a shows the confusion matrix that further quantifies the classification performance of PGCT-Net in the ONFH MRI staging task. The results demonstrate that the model achieves excellent recognition capability for the samples. In addition, the t-SNE feature visualization results in Figure 7b reveal that samples of different stages form distinct clustering structures in the two-dimensional feature space: samples within each category are closely clustered, indicating that the features extracted by the model present favorable intra-class consistency; the boundaries between different categories are clearly distinguishable, which further validates the effectiveness of the model in feature representation and category discrimination.

Table 3 presents the per-class diagnostic performance of the proposed PGCT-Net for each JIC subtype of ONFH on the independent test set. The model achieved stable and excellent diagnostic performance across all four subtypes, with sensitivity (recall) exceeding 92% for all subtypes, which fully demonstrates the balanced classification capability of the model for ONFH with different collapse risks. In particular, for Type C2—the subtype with the highest femoral head collapse risk (up to 84.8%) and the core population for early clinical intervention—PGCT-Net achieved a sensitivity of 96.15%, indicating that the model can accurately identify high-risk cases with an extremely low missed diagnosis rate, which meets the core requirements of clinical hip-preserving treatment decision-making. For Type C1, the subtype with subtle millimeter-level imaging differences from Type C2 and a high clinical misclassification rate, the model still maintained a sensitivity of 92.06%. This performance benefit comes from the hybrid CNN-Swin Transformer architecture, which simultaneously captures the fine-grained local features of lesion boundaries and the global spatial relationship between the necrotic lesion and the acetabular weight-bearing area. The overall macro-averaged sensitivity, F1-score and AUC reached 94.69%, 94.14% and 93.96% respectively, further verifying that the model maintains excellent classification performance for both majority and minority classes on the imbalanced medical dataset, without performance bias towards the majority class.

5.3.2. Experimental Results on the Dataset BT

To verify the cross-task generalization capability of the proposed PGCT-Net architecture, this study employed the BT dataset for model training and testing. The number of training epochs was set to 100, and all experimental parameters were kept consistent to objectively evaluate the comprehensive performance of the model across different medical image classification tasks. The corresponding experimental results are shown in Figures 8 and 9, and the detailed per-class diagnostic performance metrics are summarized in Table 4.

Table 4. Per-class diagnostic performance metrics of PGCT-Net on the BT test set (mean $\pm$ 95% CI).

ONFH JIC Subtype	Recall (%)	Precision (%)	F1 (%)	AUC (%)
Meningioma	94.02 $\pm$ 2.58	95.90 $\pm$ 1.83	94.93 $\pm$ 1.42	98.63 $\pm$ 0.78
Glioma	92.04 $\pm$ 2.94	97.85 $\pm$ 1.65	94.83 $\pm$ 1.58	99.10 $\pm$ 0.64
Pituitary	99.98 $\pm$ 0.02	94.32 $\pm$ 1.98	97.07 $\pm$ 1.12	99.32 $\pm$ 0.52
Macro-averaged	95.35 $\pm$ 1.79	96.02 $\pm$ 1.27	96.02 $\pm$ 1.27	95.61 $\pm$ 1.04

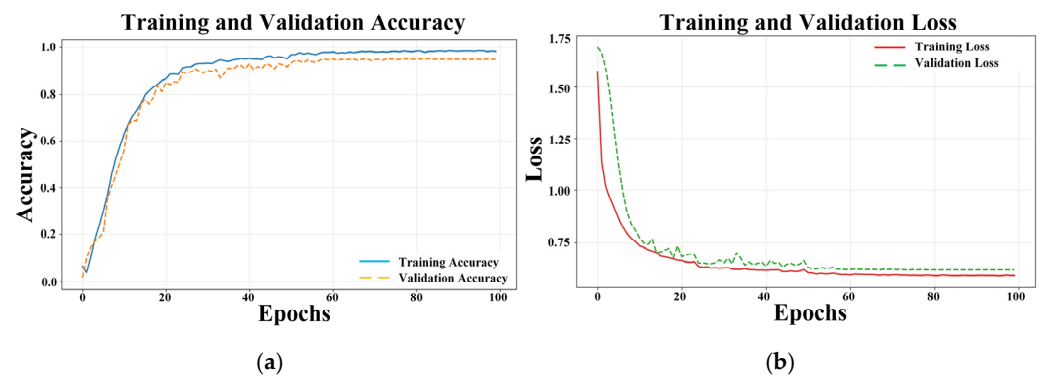


Figure 8. Convergence curves of the PGCT-Net on the dataset BT. (a) Accuracy curve; (b) loss curve.

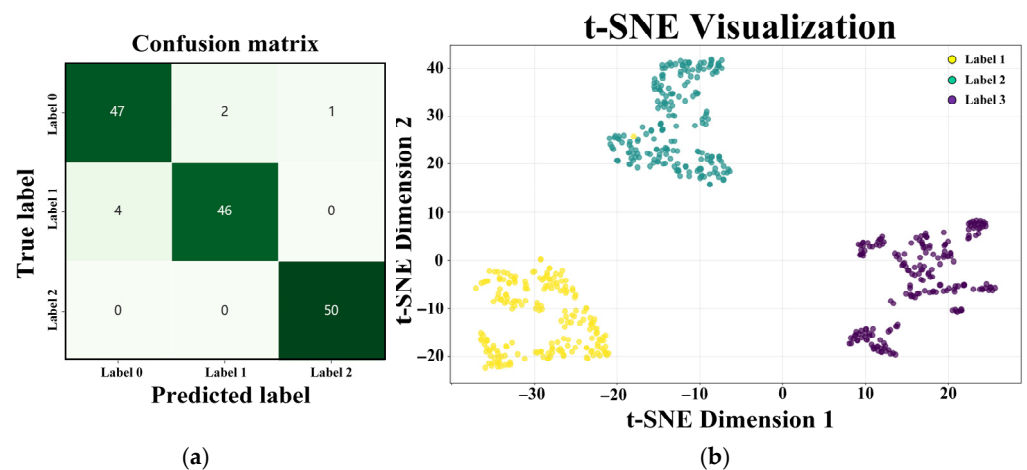


Figure 9. Result visualization of the PGCT-Net on the dataset BT. (a) Confusion matrix; (b) t-SNE visualization result.

Figure 8a shows the training and validation accuracy curves of the model. During the first 20 iterations, the training accuracy and validation accuracy increase rapidly and synchronously. The training accuracy gradually converges to nearly 100%, while the validation accuracy remains stable at 95.33%. Figure 8b presents the corresponding training and validation loss curves. Both training loss and validation loss decrease rapidly within the first 20 iterations. The training loss gradually converges to an extremely low level close to 0, and the validation loss also remains stable in a low range, which confirms the stability and convergence reliability of the model training process. The confusion matrix and t-SNE feature visualization results further validate the superior classification performance of PGCT-Net in the brain tumor MRI image classification task, as shown in Figure 9. The above experimental results demonstrate that the proposed model maintains stable and superior classification performance in both the JIC classification of osteonecrosis of the femoral head and brain tumor classification tasks. This verifies that the PGCT-Net architecture presented in this paper possesses favorable stability across different medical image classification scenarios.

Table 4 shows the per-class classification performance of PGCT-Net on the BT dataset for cross-task validation. The model achieved stable and reliable diagnostic performance across all three brain tumor subtypes, with sensitivity (recall) exceeding 92% for each category, which fully reflects the balanced recognition ability of the model for different tumor types with varied imaging features. In particular, for Pituitary tumor, PGCT-Net achieved a sensitivity of 99.98%, indicating that the model can accurately identify this typical tumor type with an extremely low missed diagnosis rate. For Meningioma and Glioma, two subtypes with complex lesion morphology, blurred boundaries

and high clinical differentiation difficulty, the model still maintained high sensitivity of 94.02% and 92.04% respectively. This stable cross-task performance benefits from the prior knowledge-guided CNN-Swin Transformer hybrid architecture, which can effectively capture local textural details and global semantic information of lesions in different anatomical parts. The overall macro-averaged sensitivity, F1-score and AUC reached 95.35%, 96.02% and 95.61% respectively, further verifying that the model maintains excellent and unbiased classification performance on different imbalanced medical image datasets.

5.4. Ablation Experiments

To verify the effectiveness of the hybrid model and prior knowledge guidance, this study conducts ablation experiments based on the ONFH MRI image dataset. Four groups of comparative models are set up in the experiment, and the specific grouping is as follows: Group 1 (CNN), Group 2 (Swin Transformer), Group 3 (Hybrid model of CNN and Swin Transformer), and Group 4 (PGCT-Net). The experimental results are shown below.

Table 5 presents the ablation experiment results of different experimental groups. It can be seen from the table that the hybrid model of CNN and Swin Transformer outperforms the single models in all metrics, with ACC, F1 and AUC values of 85.17%, 84.96% and 84.79%, respectively. This is because the hybrid model collaboratively extracts local texture features and global semantic features through a cascaded dual-branch structure, which can characterize lesion features more comprehensively and enhance the discriminative ability of the model. On the basis of the hybrid model, the introduction of the prior mask mechanism further improves the model performance, with ACC, F1 and AUC values of 94.38%, 94.15% and 93.97%, respectively, achieving the optimal classification performance. The experimental results demonstrate that prior knowledge guidance can effectively focus on the lesion region, reduce redundant information, and improve the classification performance of the model.

Table 5. Ablation study results on the JLU-ONFH test set.

Experimental Group	ACC (%)	F1 (%)	AUC (%)
Group 1	51.42 ± 4.21	51.03 ± 4.18	50.86 ± 3.79
Group 2	72.35 ± 3.17	72.14 ± 2.98	71.92 ± 2.76
Group 3	85.17 ± 1.98	84.96 ± 2.07	84.79 ± 1.46
Group 4	94.38 ± 1.03	94.15 ± 0.67	93.97 ± 0.91

Figures 10 and 11 show the t-SNE feature visualization results and confusion matrices of different experimental groups. It can be seen from the figures that the features extracted by the CNN model present severe inter-class overlap. Compared with CNN, the Swin Transformer model has improved inter-class discriminability, but still exhibits obvious intra-class dispersion and inter-class overlap. The hybrid model of CNN and Swin Transformer achieves the preliminary separation of major categories, yet there is still partial overlap among features of similar grades. In contrast, the feature distribution of the PGCT-Net proposed in this paper shows distinct inter-class separation and compact intra-class aggregation. Meanwhile, the number of misclassified samples in its confusion matrix is relatively low, with the highest diagnostic accuracy, which fully proves the excellent feature extraction and category discrimination abilities of the proposed model.

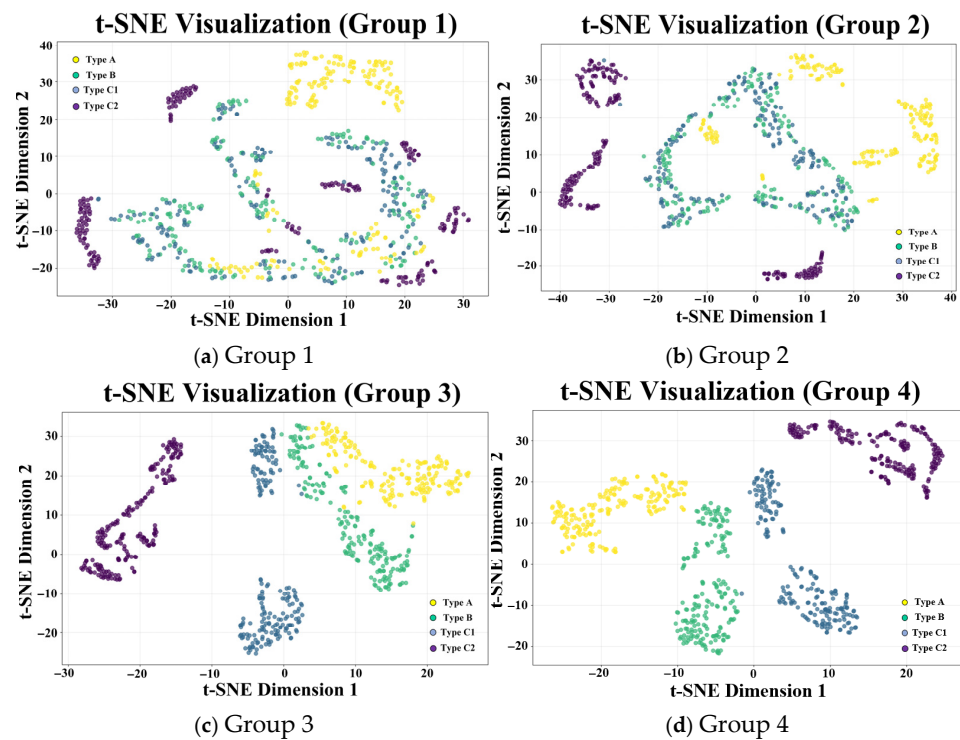


Figure 10. t-SNE visualization results of different experimental groups.

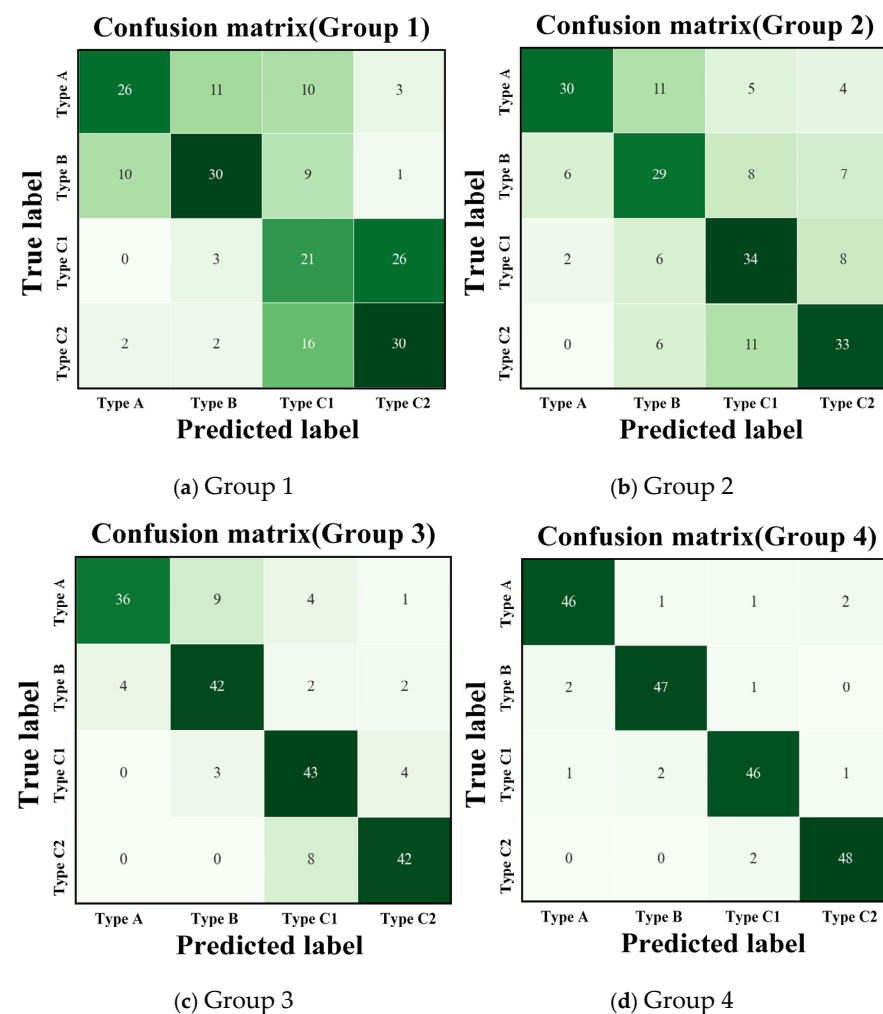


Figure 11. Confusion matrices of different experimental groups.

5.5. Effectiveness Analysis of Prior Knowledge Guidance Based on Grad-CAM

This paper adopts Gradient-weighted Class Activation Mapping (Grad-CAM) to conduct qualitative visualization analysis of the model's attention distribution, so as to verify the effectiveness of the prior knowledge guidance mechanism. By presenting the attention activation regions corresponding to classification decisions, this paper analyzes whether the mask learning mechanism can guide the model to focus on clinical pathological areas and whether the attention distribution can accurately cover lesion regions. The relevant results are illustrated in Figure 12.

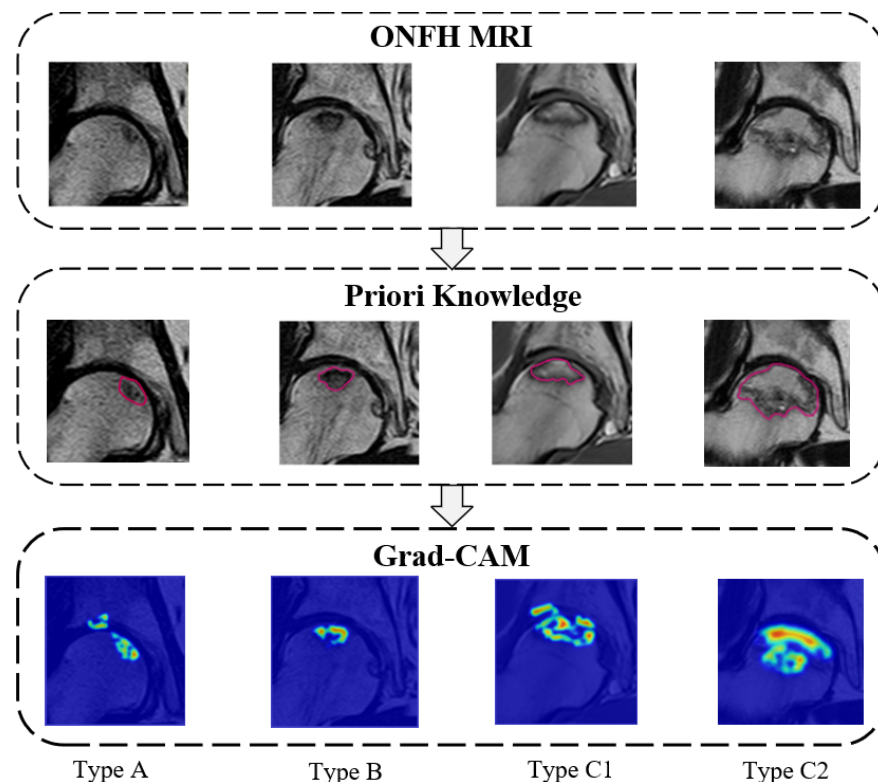


Figure 12. Lesion-focused Grad-CAM visualization guided by prior knowledge.

As shown in Figure 12, the highlighted attention regions for model decision-making are spatially consistent with the lesion mask regions annotated by clinicians. The activation regions are mainly concentrated on the core diagnostic areas including the necrotic core, sclerotic band, cystic degeneration and bone marrow edema. The model obviously suppresses invalid activation in irrelevant background such as the acetabulum, muscles and adipose tissues, and reduces redundant attention on normal bone tissue. Even for early tiny lesions, the attention regions of the model match the lesion boundaries and core scope well, without obvious dispersion or deviation.

Through end-to-end mask constraint and joint optimization, clinical prior knowledge is deeply embedded into the feature learning process, which reasonably calibrates the model's learning objectives. The model mainly makes classification decisions based on core lesion features recognized clinically, instead of relying on background pseudo-features. Its attention distribution is consistent with clinicians' diagnostic logic of focusing on lesions, which improves the interpretability of the model and reduces the black-box characteristic of deep learning models to a certain extent. This provides intuitive support for the clinical practicability of the model [5,12].

5.6. Comparative Experiments

To comprehensively verify the overall performance and generalization capability of the proposed PGCT-Net in medical image classification tasks, nine state-of-the-art mainstream medical image classification methods are selected for comparative analysis, including R2C-GAN [38], ThoraX-PriorNet [39], ARiViT [40], Residual-BRNet [41], Integrated Attention-Guided CNN [42], CCF-GNN [43], CDAE-R [44], GPTNeXt [45] and CNN-ViT [46]. All comparative models are strictly reimplemented in accordance with the experimental procedures described in the original literature. Comparative experiments are performed on the JLU-ONFH dataset and the BT dataset. All baseline models adopt the same complete experimental workflow as PGCT-Net. Unified arrangements are implemented for dataset split, image preprocessing, data augmentation strategy, optimizer configuration, learning rate scheduler, batch size, early stopping rule, and model selection criterion. The optimal model is determined based on the highest macro average AUC on the validation set. The absolute fairness and reproducibility of comparative experiments are therefore guaranteed.

To ensure the reliability and stability of the experimental results, all experiments are repeated independently 10 times, and the average values of various indicators are taken as the final results to minimize the interference of random factors on the experimental outcomes during the training process. The experimental results are presented as follows.

The classification performance comparison results between each comparative model and the proposed PGCT-Net on the dataset JLU-ONFH are depicted in Figure 13. R2C-GAN expands the feature space via sample generation, yet targeted learning of fine-grained lesion textures is absent. The capacity for discriminative feature extraction is compromised, resulting in suboptimal classification accuracy [38]. ThoraX-PriorNet incorporates medical prior knowledge to enhance the recognition of positive samples, achieving an AUC of 93.20%. However, prior constraints are not embedded into the feature learning pipeline, leading to insufficient classification balance and an F1-score of only 87.80% [39]. ARiViT performs global feature modeling through the self-attention mechanism, with ACC and F1-score reaching 93.05% and 92.50%, respectively. This model lacks the inductive bias for local texture representation, exhibits poor capability in extracting detailed features of ambiguous lesions, and demonstrates limited adaptability to small-sample data. A substantial performance gap remains between this approach and the model proposed herein [40].

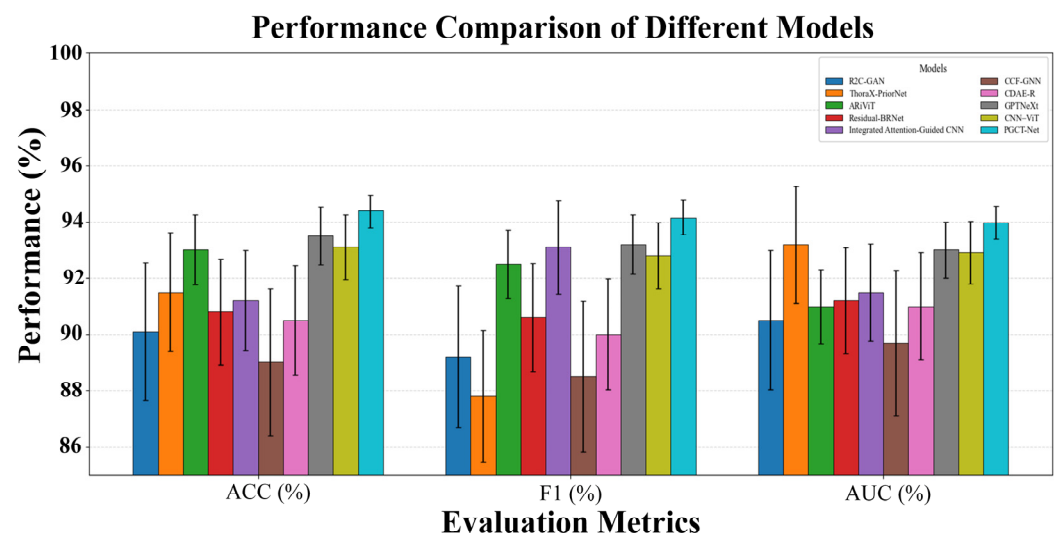


Figure 13. Performance comparison of different models on the JLU-ONFH dataset.

Residual-BRNet enables stable extraction of local lesion features via residual connections, with all three evaluation metrics stabilized at approximately 91%. Restricted by the finite receptive fields of convolutional kernels, long-range dependencies between lesions and global structural information cannot be effectively modeled [41]. Integrated Attention-Guided CNN leverages an attention mechanism to focus on lesion regions, but the granularity of feature localization is coarse. Weak noise robustness and generalization capability are observed, accompanied by a low overall accuracy of 91.20% [42]. CCF-GNN models feature correlations based on graph structures, yet it is poorly compatible with the two-dimensional spatial features of MRI images, failing to effectively capture the spatial distribution and morphological characteristics of lesions [43]. CDAE-R accomplishes feature denoising and reconstruction using autoencoders. With unsupervised reconstruction as the learning objective, this framework is decoupled from the classification task, impeding the mining of high-level discriminative lesion features [44]. GPTNeXt extracts multi-domain features by combining large-kernel convolution and Transformer architectures, but the multi-scale feature fusion strategy is oversimplified, making it difficult to accommodate the high variability in lesion size and location [45]. CNN-ViT integrates the complementary advantages of convolution and Transformer modules. Nevertheless, the feature interaction mechanism is underdeveloped, and the fusion of local details and global context is inadequate. Low feature utilization efficiency is induced, and thus its performance is inferior to that of the proposed model [46].

As quantitatively illustrated in Table 6, the proposed PGCT-Net achieves the optimal performance in ACC ($94.38 \pm 1.03\%$), F1-score ($94.15 \pm 0.67\%$), and AUC ($93.97 \pm 0.91\%$) among all baseline models. Crucially, PGCT-Net exhibits the smallest standard deviation across all metrics, proving superior stability and reproducibility compared with all competing methods. The consistent performance superiority over direct ablation baselines, prior-guided models, and state-of-the-art approaches further validates the effectiveness of the mask-guided prior knowledge mechanism and the hybrid CNN-Swin architecture designed in this work.

Table 6. Comprehensive performance comparison of all models on the JLU-ONFH dataset.

Model Category	Model Name	ACC (%)	F1 (%)	AUC (%)
Core Direct Baselines	CNN	51.42 ± 4.21	51.03 ± 4.18	50.86 ± 3.79
	Swin Transformer	72.35 ± 3.17	72.14 ± 2.98	71.92 ± 2.76
	CNN-Swin Hybrid	85.17 ± 1.98	84.96 ± 2.07	84.79 ± 1.46
Prior & Attention-Guided SOTA Models	ThoraX-PriorNet [39]	91.32 ± 1.74	87.80 ± 2.11	93.20 ± 1.53
	Integrated Attention-Guided CNN [42]	91.20 ± 1.86	90.85 ± 1.93	90.62 ± 1.74
Other Mainstream SOTA Models	R2C-GAN [38]	88.45 ± 2.05	87.92 ± 2.12	87.54 ± 1.98
	ARiViT [40]	93.05 ± 1.52	92.50 ± 1.64	92.17 ± 1.58
	Residual-BRNet [41]	91.04 ± 1.79	90.76 ± 1.85	90.53 ± 1.69
	CCF-GNN [43]	89.37 ± 1.92	89.02 ± 2.01	88.76 ± 1.83
	CDAE-R [44]	92.14 ± 1.67	91.89 ± 1.73	91.57 ± 1.65
	GPTNeXt [45]	90.78 ± 1.82	90.43 ± 1.88	90.11 ± 1.76
	CNN-ViT [46]	92.56 ± 1.59	92.17 ± 1.66	91.89 ± 1.61
Proposed Model	PGCT-Net	94.38 ± 1.03	94.15 ± 0.67	93.97 ± 0.91

The classification performance comparison results between each comparative model and the proposed PGCT-Net on the dataset BT are illustrated in Figure 14. R2C-GAN shows a slight performance improvement on the BT dataset but remains at a below-average level overall. Although this model enriches feature representation through a generative mechanism, it presents poor adaptability to medical image features with distinct distributions and fails to effectively handle variations in lesion morphology, leading to limited

promotion in classification accuracy [38]. ThoraX-PriorNet achieves an ACC of 92.80%, yet its AUC is only 91.00%. The issue of unbalanced performance is further intensified with obvious performance degradation, and its robustness is far inferior to that of the proposed PGCT-Net [39].

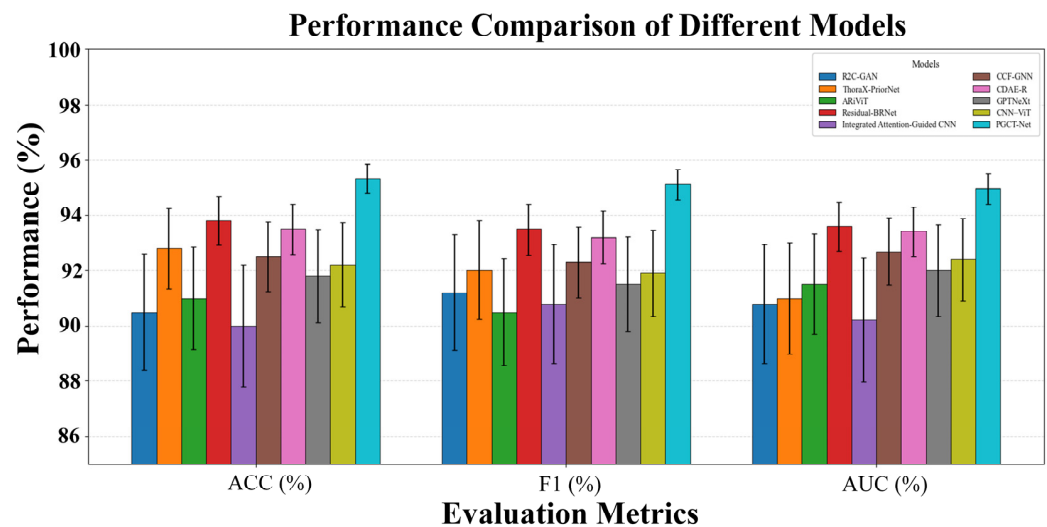


Figure 14. Performance comparison of different models on the BT dataset.

ARiViT exhibits a remarkable performance drop, as it is sensitive to changes in data distribution and lacks sufficient generalization capability [40]. Residual-BRNet obtains a substantial performance improvement on the BT dataset but still cannot break through the inherent limitation of convolutional structures in modeling global information, thus underperforming PGCT-Net [41]. Integrated Attention-Guided CNN suffers from a drastic performance decline, with significantly weakened feature extraction ability and unsatisfactory overall performance [42]. CCF-GNN has better adaptability to the spatial correlation features of lesions in the BT dataset owing to its graph structure modeling capacity, yet it still has obvious deficiencies in extracting low-level detailed features of images, resulting in a notable gap in overall performance compared with the proposed model [43].

CDAE-R achieves all three evaluation metrics above 93%. Its denoising and reconstruction based on an autoencoder possesses strong robustness in extracting lesion features, but, constrained by the upper limit of feature expression of the encoder, it fails to deeply mine multi-scale and global features of lesions, so its performance is still weaker than PGCT-Net [44]. The hybrid structure of GPTNeXt is sensitive to variations in data distribution, and its fusion mechanism for local and global features cannot function effectively. With insufficient generalization ability, the performance gap between GPTNeXt and the proposed model is further widened [45]. Although the CNN-Transformer hybrid architecture of CNN-ViT has certain feature adaptability, it cannot realize dynamic optimization for lesion features of different datasets. The depth and effectiveness of feature fusion are insufficient, leading to poorer performance than PGCT-Net [46].

In accordance with the quantitative results listed in Table 7, the proposed PGCT-Net exhibits promising classification performance on the BT dataset and achieves competitive ACC ($95.33 \pm 0.98\%$), F1-score ($95.61 \pm 0.87\%$), and AUC ($99.02 \pm 0.76\%$) among all comparative models. Furthermore, PGCT-Net produces lower standard deviations for all evaluation metrics than other baseline models.

Table 7. Comprehensive performance comparison of all models on the BT dataset.

Model Category	Model Name	ACC (%)	F1 (%)	AUC (%)
Core Direct Baselines	CNN	58.26 ± 3.87	57.93 ± 3.92	57.68 ± 3.64
	Swin Transformer	76.43 ± 2.95	76.18 ± 3.02	75.84 ± 2.83
	CNN-Swin Hybrid	87.25 ± 1.86	86.97 ± 1.94	86.72 ± 1.58
Prior & Attention-Guided SOTA Models	ThoraX-PriorNet [39]	92.80 ± 1.69	91.05 ± 1.87	91.00 ± 1.75
	Integrated Attention-Guided CNN [42]	89.64 ± 1.91	89.27 ± 1.98	88.93 ± 1.86
Other Mainstream SOTA Models	R2C-GAN [38]	90.12 ± 1.88	89.76 ± 1.95	89.42 ± 1.82
	ARiViT [40]	91.76 ± 1.73	91.34 ± 1.81	90.98 ± 1.77
	Residual-BRNet [41]	92.05 ± 1.68	91.72 ± 1.75	91.38 ± 1.69
	CCF-GNN [43]	92.34 ± 1.62	92.01 ± 1.69	91.67 ± 1.64
	CDAE-R [44]	93.12 ± 1.58	92.87 ± 1.64	92.54 ± 1.61
	GPTNeXt [45]	90.87 ± 1.84	90.53 ± 1.90	90.19 ± 1.83
	CNN-ViT [46]	93.47 ± 1.54	93.12 ± 1.60	92.78 ± 1.57
Proposed Model	PGCT-Net	95.33 ± 0.98	95.61 ± 0.87	99.02 ± 0.76

Combined with the experimental results of the two datasets, the following conclusion can be drawn: with changes in the anatomical regions and feature distributions of medical images, most comparative models exhibit varying degrees of performance fluctuation and accuracy degradation. In contrast, the proposed PGCT-Net achieves a smaller range of performance variation in different medical image classification tasks and stably maintains high classification accuracy. This demonstrates that the hybrid architecture integrated with prior knowledge guidance presents stronger adaptability to diverse medical image feature distributions, and the developed algorithm possesses favorable stability in medical image classification tasks.

6. Conclusions

Aiming at the core challenges in the clinical diagnosis of JIC classification for osteonecrosis of the ONFH, including large subjective errors, difficulties in identifying early-stage lesions, insufficient feature adaptability of existing AI models, and inadequate model reliability, this paper proposes a Prior Knowledge-Guided CNN-Swin Transformer Hybrid Network, namely PGCT-Net, for the accurate classification and diagnosis of ONFH using MRI images. The model adopts a multi-stage cascaded architecture, which integrates the strong capability of CNN in capturing local fine-grained lesion features and the advantage of Swin Transformer in modeling global spatial semantic features. Meanwhile, it introduces a clinical lesion mask-guided learning mechanism to deeply embed clinical diagnostic prior knowledge into the whole process of feature learning and decision-making of the model, realizing the deep coupling of local and global features, and enhancing the visual interpretability and reliability of model decision-making. The method proposed in this paper can provide an efficient, reliable and visually interpretable AI-assisted diagnostic tool for accurate JIC classification of ONFH and individualized hip-preserving treatment.

Author Contributions: Conceptualization, C.Y. and M.W.; Methodology, C.Y. and M.W.; Software, C.Y.; Validation, C.Y. and M.W.; Formal analysis, C.Y.; Investigation, C.Y.; Data curation, C.Y.; Writing—original draft, C.Y.; Writing—review & editing, C.Y. and M.W.; Visualization, C.Y.; Supervision, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in the following publicly accessible repositories. The Jilin University Osteonecrosis of the Femoral Head (JLU-ONFH) dataset is available at <https://github.com/sole12343/JLU-ONFH-Dataset> (accessed on 13 April 2026), and the Brain Tumor (BT) dataset is available at https://figshare.com/articles/dataset/brain_tumor_dataset/1512427 (accessed on 13 April 2026).

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

References

- Shen, X.; He, Z.; Shi, Y.; Liu, T.; Yang, Y.; Luo, J.; Tang, X.; Chen, B.; Xu, S.; Zhou, Y. Development and validation of an automated classification system for osteonecrosis of the femoral head using deep learning approach: A multicenter study. *J. Arthroplast.* **2024**, *39*, 379–386.e2. [\[CrossRef\]](#)
- Zhao, D.; Zhang, F.; Wang, B.; Liu, B.; Li, L.; Kim, S.-Y.; Goodman, S.B.; Hernigou, P.; Cui, Q.; Lineaweaver, W.C. Guidelines for clinical diagnosis and treatment of osteonecrosis of the femoral head in adults (2019 version). *J. Orthop. Transl.* **2020**, *21*, 100–110. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tan, B.; Li, W.; Zeng, P.; Guo, H.; Huang, Z.; Fu, F.; Gao, H.; Wang, R.; Chen, W. Epidemiological study based on China osteonecrosis of the femoral head database. *Orthop. Surg.* **2021**, *13*, 153–160. [\[CrossRef\]](#)
- Liu, W.; Yue, J.; Guo, X.; Wang, R.; Fu, H. Epidemiological investigation and diagnostic analysis of osteonecrosis of the femoral head in three northeastern provinces of China. *J. Orthop. Surg. Res.* **2024**, *19*, 292. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kuroda, Y.; Tanaka, T.; Miyagawa, T.; Kawai, T.; Goto, K.; Tanaka, S.; Matsuda, S.; Akiyama, H. Classification of osteonecrosis of the femoral head: Who should have surgery? *Bone Jt. Res.* **2019**, *8*, 451–458. [\[CrossRef\]](#)
- Kalekar, T.; Bothara, K.; Lamghare, P. The Application of Magnetic Resonance Imaging in the Early and Accurate Diagnosis of Hip Joint Avascular Necrosis. *Cureus* **2024**, *16*, e68888. [\[CrossRef\]](#)
- Ge, H.; Wang, Z.; Zhang, J. X-ray, digital tomographic fusion, CT, and MRI in early ischemic necrosis of the femoral head. *Medicine* **2024**, *103*, e36281. [\[CrossRef\]](#)
- Parikh, R.R.; Mirzaei, A.; Butler, M.E.; Restrepo, D.J.; Guarin Perez, S.F.; Brandt, S.; Swartz, G.; Katanbaf, R.; Goodman, S.B.; Mont, M.A. Diagnosis and Treatment of Nontraumatic Osteonecrosis of the Femoral Head: A Systematic Review and Meta-Analyses for the ARCO Clinical Practice Guideline Development Workgroup. *Med. Sci.* **2026**, *14*, 107. [\[CrossRef\]](#)
- Yoon, B.-H.; Mont, M.A.; Koo, K.-H.; Chen, C.-H.; Cheng, E.Y.; Cui, Q.; Drescher, W.; Gangji, V.; Goodman, S.B.; Ha, Y.-C. The 2019 revised version of association research circulation osseous staging system of osteonecrosis of the femoral head. *J. Arthroplast.* **2020**, *35*, 933–940. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pasko, K.B.D.; Bohlen, H.; Kim, M.S.; Ray, L.; Gallagher, B.; Tomasian, A.; Fader, R.; Wang, D. Geographic characterization of intra-articular pathologies in early-stage avascular necrosis of the hip. *BMC Musculoskelet. Disord.* **2026**, *27*, 202. [\[CrossRef\]](#)
- Sugano, N.; Atsumi, T.; Ohzono, K.; Kubo, T.; Hotokebuchi, T.; Takaoka, K. The 2001 revised criteria for diagnosis, classification, and staging of idiopathic osteonecrosis of the femoral head. *J. Orthop. Sci.* **2002**, *7*, 601–605. [\[CrossRef\]](#) [\[PubMed\]](#)
- Takashima, K.; Sakai, T.; Hamada, H.; Takao, M.; Sugano, N. Which classification system is most useful for classifying osteonecrosis of the femoral head? *Clin. Orthop. Relat. Res.* **2018**, *476*, 1240–1249. [\[CrossRef\]](#) [\[PubMed\]](#)
- Nakamura, J.; Kishida, S.; Harada, Y.; Iida, S.; Oinuma, K.; Yamamoto, S.; Nakajima, T.; Takazawa, M.; Shigemura, T.; Ohtori, S. Inter-observer and intra-observer reliabilities of the Japanese Ministry of Health, Labor and Welfare type classification system for osteonecrosis of the femoral head. *Mod. Rheumatol.* **2011**, *21*, 488–494. [\[CrossRef\]](#) [\[PubMed\]](#)
- Alkhatatbeh, T.; Alkhatatbeh, A.; Li, X.; Wang, W. A single sequence MRI-based deep learning radiomics model in the diagnosis of early osteonecrosis of femoral head. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1471692. [\[CrossRef\]](#)
- Li, X.; Lv, S.; Tong, C.; Qin, Y.; Liang, C.; Ma, Y.; Li, M.; Luo, H.; Yin, S. MsgeCNN: Multiscale geometric embedded convolutional neural network for ONFH segmentation and grading. *Med. Phys.* **2023**, *50*, 3788–3800. [\[CrossRef\]](#)
- Klontzas, M.E.; Vassalou, E.E.; Spanakis, K.; Meurer, F.; Woertler, K.; Zibis, A.; Marias, K.; Karantanas, A.H. Deep learning enables the differentiation between early and late stages of hip avascular necrosis. *Eur. Radiol.* **2024**, *34*, 1179–1186. [\[CrossRef\]](#)
- Zarch, N.E.; Bagher-Ebadian, H.; Alhanai, T.; Ghassemi, M.M. Glog-csunet: Enhancing vision transformers with adaptable radiomic features for medical image segmentation. In *Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.
- Deng, C.; Qin, X. MINTFormer: Multi-scale information aggregation with CSWin vision transformer for medical image segmentation. *Appl. Sci.* **2025**, *15*, 8626. [\[CrossRef\]](#)
- Rahman, M.I.; Rahman, A. AACNN-ViT: Adaptive Attention-Augmented Convolutional and Vision Transformer Fusion for Lung Cancer Detection. *J. Imaging* **2026**, *12*, 62. [\[CrossRef\]](#)

20. Uemura, K.; Takashima, K.; Otake, Y.; Li, G.; Mae, H.; Okada, S.; Hamada, H.; Sugano, N. Development of a Deep Learning Model for the Volumetric Assessment of Osteonecrosis of the Femoral Head on Three-Dimensional Magnetic Resonance Imaging. *J. Arthroplast.* **2025**, *40*, S160–S166.e1. [\[CrossRef\]](#)
21. Liang, C.; Ma, Y.; Li, X.; Qin, Y.; Li, M.; Tong, C.; Xu, X.; Yu, J.; Wang, R.; Lv, S. Aiding Diagnosis and Classifying of Early Stage Osteonecrosis of the Femoral Head with Convolutional Neural Network Based on Magnetic Resonance Imaging. *Indian J. Orthop.* **2025**, *59*, 121–127. [\[CrossRef\]](#)
22. Lee, H.H.; Jeong, J.; Shin, T.; Chae, D.-S. Classification of Osteonecrosis of the Femoral Head Stage on Radiographic Images Using Deep Learning Techniques. *Bioengineering* **2025**, *12*, 1319. [\[CrossRef\]](#)
23. Haupt, M.; Maurer, M.H.; Thomas, R.P. Explainable Artificial Intelligence in Radiological Cardiovascular Imaging—A Systematic Review. *Diagnostics* **2025**, *15*, 1399. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9. [\[CrossRef\]](#)
25. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [\[CrossRef\]](#)
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778. [\[CrossRef\]](#)
27. 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. [\[CrossRef\]](#)
28. Ferdous, M.; Mahmud, S.; Shimul, M.E.Z. MedNet: A lightweight attention-augmented CNN for medical image classification. *Sci. Rep.* **2025**, *15*, 41936. [\[CrossRef\]](#)
29. Li, J.; Goh, W.; Jhanjhi, N.Z. A lightweight CNN for colon cancer tissue classification and visualization. *Front. Oncol.* **2025**, *15*, 1659010. [\[CrossRef\]](#)
30. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [\[CrossRef\]](#)
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022. [\[CrossRef\]](#)
32. Chen, G.; Zhang, H.; Zeng, J.; Cai, Y.; Huang, D.; Chen, Y.; Li, P.; Zheng, Y. Benchmark-Driven Clinical Decision Framework for Multi-Class Middle Ear Disease Diagnosis: Superiority of Swin Transformer in Accuracy and Stability. *Diagnostics* **2026**, *16*, 482. [\[CrossRef\]](#)
33. Xu, Y.; Ying, W.; Du, Y.; Wang, Q.; Zheng, J.; Yang, Z. Swin Transformer-based intelligent fracture classification and radiographic assessment to support clinical decision-making. *BMC Med. Inform. Decis. Mak.* **2026**, *26*, 42. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Wang, X.; Wang, G.; Li, L.; Zou, H.; Cui, J. MFF-ClassificationNet: CNN-Transformer Hybrid with Multi-Feature Fusion for Breast Cancer Histopathology Classification. *Biosensors* **2025**, *15*, 718. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Qezelbash-Chamak, J.; Hicklin, K. A hybrid learnable fusion of convnext and swin transformer for optimized image classification. *IoT* **2025**, *6*, 30. [\[CrossRef\]](#)
36. Shen, X.; Luo, J.; Tang, X.; Chen, B.; Qin, Y.; Zhou, Y.; Xiao, J. Deep learning approach for diagnosing early osteonecrosis of the femoral head based on magnetic resonance imaging. *J. Arthroplast.* **2023**, *38*, 2044–2050. [\[CrossRef\]](#)
37. Cheng, J.; Huang, W.; Cao, S.; Yang, R.; Yang, W.; Yun, Z.; Wang, Z.; Feng, Q. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE* **2015**, *10*, e0140381. [\[CrossRef\]](#)
38. Ahishali, M.; Degerli, A.; Kiranyaz, S.; Hamid, T.; Mazhar, R.; Gabbouj, M. R2C-GAN: Restore-to-Classify Generative Adversarial Networks for blind X-ray restoration and COVID-19 classification. *Pattern Recognit.* **2024**, *156*, 110765. [\[CrossRef\]](#)
39. Hossain, M.I.; Zunaed, M.; Ahmed, M.K.; Hossain, S.J.; Hasan, A.; Hasan, T. ThoraX-PriorNet: A novel attention-based architecture using anatomical prior probability maps for thoracic disease classification. *IEEE Access* **2023**, *12*, 3256–3273. [\[CrossRef\]](#)
40. Hameed, M.; Zameer, A.; Khan, S.H.; Raja, M.A.Z. ARiViT: Attention-based residual-integrated vision transformer for noisy brain medical image classification. *Eur. Phys. J. Plus* **2024**, *139*, 440. [\[CrossRef\]](#)
41. Alshemaimri, B.K. Novel deep CNNs explore regions, boundaries, and residual learning for COVID-19 infection analysis in lung CT. *Tomography* **2024**, *10*, 1205–1221. [\[CrossRef\]](#)
42. Öksüz, C.; Urhan, O.; Güllü, M.K. An integrated convolutional neural network with attention guidance for improved performance of medical image classification. *Neural Comput. Appl.* **2024**, *36*, 2067–2099. [\[CrossRef\]](#)
43. Wang, H.; Huang, G.; Zhao, Z.; Cheng, L.; Juncker-Jensen, A.; Nagy, M.L.; Lu, X.; Zhang, X.; Chen, D.Z. CCF-GNN: A unified model aggregating appearance, microenvironment, and topology for pathology image classification. *IEEE Trans. Med. Imaging* **2023**, *42*, 3179–3193. [\[CrossRef\]](#) [\[PubMed\]](#)

44. Wang, Z. CDAE-R: Multifunctional End-to-End Model for Brain Abnormality Images Classification and Denoising. In *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*; Atlantis Press: Dordrecht, The Netherlands, 2024; pp. 930–938. [[CrossRef](#)]
45. Alotaibi, F.A.; Yagmahan, M.S.N.; Alobaid, K.A.; Jari, M.; Goktas, O.F.; Baygin, M.; Tuncer, T.; Dogan, S. GPTNeXt: Biomedical Image Classification Investigations. *Diagnostics* **2026**, *16*, 581. [[CrossRef](#)] [[PubMed](#)]
46. Almughamisi, N.; Abosamra, G.; Albar, A.; Saleh, M. Anatomy-Guided Hybrid CNN–ViT Model with Neuro-Symbolic Reasoning for Early Diagnosis of Thoracic Diseases Multilabel. *Diagnostics* **2026**, *16*, 159. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.