Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

Research on a Dual-Branch Deep Network Based on an Improved CBAM Attention Mechanism and Interaction-Aware Gated Fusion for Auxiliary Diagnosis of Musculoskeletal Injuries

Wenjie Zhang [1], Guangping Zhuo [1, *], Guanghua Zhang [2], Fei Ma [2]

[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong Shanxi, 030619, China
[2] Department of Computer Science and Technology, Taiyuan University, Taiyuan Shanxi, 030032, China

*** Corresponding author:** Guangping Zhuo

Abstract: To address the strong subjectivity of conventional diagnostic methods for musculoskeletal injuries, the insufficient fusion of heterogeneous multimodal features in existing gait analysis approaches, and their weak ability to focus on critical lesion regions, this paper proposes a novel dual-branch deep learning architecture for the auxiliary diagnosis of musculoskeletal injuries. The model adopts a dual-branch structure that independently extracts the spatial texture of plantar pressure distribution heatmaps and the nonlinear features of quantified gait parameters. At the architectural level, a Dynamic Cross-Channel and Multi-Scale Spatial Fusion Attention mechanism (DCMSF-CBAM) is innovatively constructed to achieve dynamic multi-path fusion along both the spatial and channel dimensions, which markedly strengthens the model's ability to focus on critical lesion regions such as the heel and the arch. In addition, an Interaction-Aware Gated Fusion module (IAGF) is designed, in which a dynamic gating mechanism driven by interaction features adaptively and weightedly integrates visual representations with quantified numerical values, effectively overcoming the information redundancy and loss caused by conventional simple concatenation. Validated on a four-class clinical gait dataset covering healthy subjects, knee osteoarthritis, anterior cruciate ligament injury, and meniscus injury, the proposed model achieves a classification accuracy of 0.9476, with precision, recall, and F1-score of 0.9456, 0.9557, and 0.9500, respectively, under the macro-averaging strategy. The results indicate that the proposed method can provide a reference for the early screening and auxiliary diagnosis of musculoskeletal injuries.

Keywords: Musculoskeletal Injury; Gait Analysis; Deep Learning; Attention Mechanism; Gated Fusion.

1. Introduction

Musculoskeletal injuries, such as knee osteoarthritis (KOA), anterior cruciate ligament (ACL) injury, and meniscus injury, are common diseases that severely affect the quality of life of middle-aged and elderly populations worldwide. As global population ageing accelerates, the incidence of such diseases continues to rise; the resulting pain, limited joint function, and gait abnormalities not only significantly reduce individual mobility but also impose a substantial medical burden[1]. Conventional diagnosis relies heavily on the subjective experience of clinicians, physical examinations, and expensive imaging modalities such as MRI, and thus suffers from strong subjectivity, long diagnostic cycles, high costs, and low detection rates for early lesions, making it difficult to meet the demands of large-scale early screening. With the integration of biomechanics and artificial intelligence, auxiliary diagnosis based on objective quantitative gait indicators has become a research hotspot[2]. Gait analysis can non-invasively and in real time capture the spatiotemporal, kinetic, and kinematic abnormalities induced by musculoskeletal injuries, thereby providing crucial evidence for early intervention.

Gait is the comprehensive external manifestation of the function of the human musculoskeletal system, and its spatiotemporal parameters (e.g., gait speed, stride length, and joint range of motion) and kinetic characteristics (e.g., plantar pressure distribution and ground reaction force) can sensitively reflect the early biomechanical alterations associated with musculoskeletal injuries. In recent years, sensors such as motion capture (MoCap) systems, wearable inertial measurement units (IMUs), surface electromyography (sEMG), and markerless Kinect systems have been widely applied to gait data acquisition, providing a rich multimodal input foundation for deep learning models. Existing studies have demonstrated that deep learning exhibits significant advantages in abnormal gait recognition. For example, Teng et al.[3] corrected the lower-limb joint angle errors of a Kinect system through the fusion of BiLSTM and linear regression, substantially improving the accuracy of a markerless gait analysis system; Bodepu et al.[4] proposed CRA-Net, which integrates two complementary attention mechanisms, namely a cross reverse attention mechanism and a cross attention mechanism, to effectively capture subtle inter-class differences in the gait of neurodegenerative diseases; Kaya et al.[5] employed a hybrid model combining a pre-trained CNN and an SVM to achieve high-accuracy detection of ataxia in patients with multiple sclerosis (PwMS) based on static plantar pressure distribution images (accuracy 95.12%); Li et al.[6] constructed a CNN-ResNet-BiLSTM model that fuses IMU and sEMG data, achieving an overall accuracy of 97.78% in abnormal gait recognition for stroke patients; Atick et al.[7] designed NDDNet, which simultaneously processes ground reaction force signals and statistical features, attaining an average classification accuracy of 96.75% for Parkinson's disease, Huntington's disease, and amyotrophic lateral sclerosis; Rios et al.[8] constructed an RGB plantar image dataset and trained six deep models for evaluation, ultimately achieving 98.013%

accuracy with Swin Transformer Tiny (Swin-T) and thereby realizing automatic screening for foot-related diseases. In addition, Kaya et al.[9] utilized an LSTM-FCN model to process markerless motion capture gait data, effectively classifying the severity grade of knee osteoarthritis (KOA) under the Kellgren-Lawrence grading system; Chakraborty et al.[10] enhanced the representational capability of a one-dimensional convolutional neural network (1D-CNN) for the abnormal gait of children with cerebral palsy through discrete wavelet transform; and Kwon et al.[11] verified the complementarity between gait data and radiographic data in KOA classification, with model performance significantly outperforming single-modality models. Collectively, these works indicate that multimodal fusion, attention mechanisms, residual connections, and graph convolution have become key pathways for improving the robustness of gait analysis.

Despite the above progress, existing deep learning-based gait diagnosis methods still face several bottlenecks. First, image modalities (e.g., plantar pressure distribution maps) and numerical modalities (e.g., gait speed, stride length, and joint range of motion) are highly heterogeneous in their features, and conventional cascade or simple concatenation fusion can hardly adapt dynamically to the feature importance of different inputs, resulting in information redundancy or loss. Second, most existing attention modules are static channel or spatial attention, lacking the capability for dynamic cross-channel interaction and multi-scale fusion, and thus cannot sufficiently focus on critical gait regions (e.g., abnormal pressure patterns around the knee joint or the temporal sequence of joint motion abnormalities). Finally, specific models for musculoskeletal injuries (e.g., KOA, ACL, and meniscus injuries) remain scarce, and most studies are limited to a single disease, making it difficult to satisfy the refined clinical requirements of multiple disease types and severity grading.

To overcome the above limitations, this paper proposes a novel dual-branch deep learning architecture dedicated to the gait-based auxiliary diagnosis of musculoskeletal injuries. The model employs Swin-T as the backbone network of the image branch and supplements it with a fully connected network that processes numerical data. On this basis, the dynamic multi-path attention of DCMSF-CBAM and the gating mechanism of IAGF are innovatively introduced, thereby focusing on critical gait regions (e.g., the heel, the arch, and the forefoot pressure concentration area) and enabling the model to flexibly balance the weights of visual texture and quantitative parameters according to input quality. This markedly improves the representational capability for injury-specific gait patterns and the cross-subject generalization performance. This study provides a reference for the early objective screening and auxiliary diagnosis of musculoskeletal diseases such as KOA, ACL, and meniscus injuries.

2. Data and Model Construction

2.1. Experimental Data

2.1.1. Description of the Experimental Data

This study employed a musculoskeletal injury dataset comprising 702 subject samples. The dataset covers clinical samples under four different conditions, including 78 healthy subjects, 481 patients with knee osteoarthritis, 81 patients with anterior cruciate ligament injury, and 62 patients with meniscus injury. The data were collected in the gait laboratory of a collaborating hospital, where a Footscan 2m HE plantar pressure distribution plate system and a Codamotion 2CX1 three-dimensional joint motion capture system were used to record the plantar gait data and lower-limb joint data of the patients. The measured data are shown in Fig. 1.

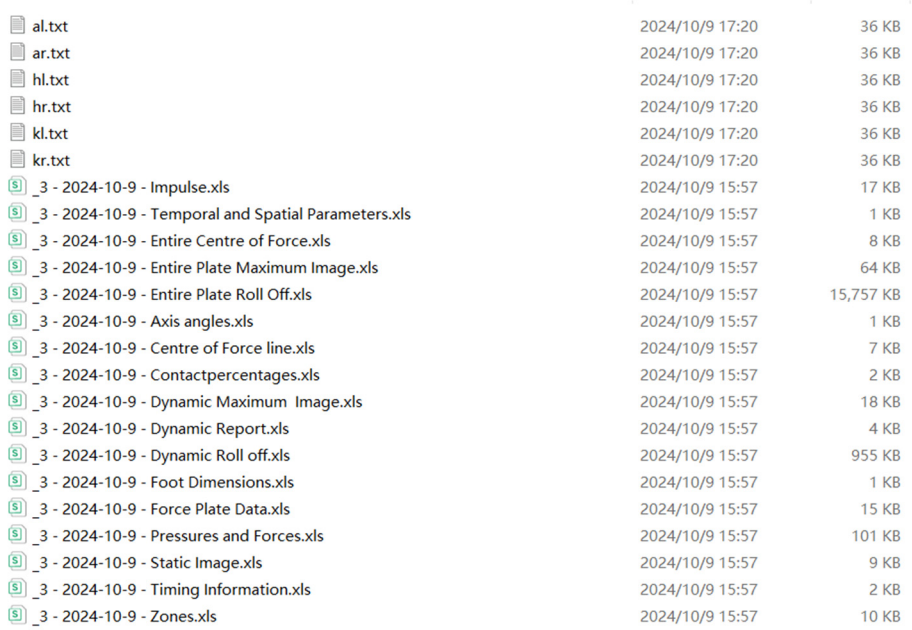


Fig 1. Gait plantar data and joint data

2.1.2. Data Preprocessing

In the preprocessing pipeline, all relevant feature columns were first screened manually; the relevant feature data selected in this paper were established after discussion with several orthopaedic experts from the collaborating hospital. After the raw data were imported from the corresponding files, the relevant feature parameters were extracted and computed using the array operations and statistical functions of the NumPy library (such as np.array, np.hstack, and max/min) and written into the corresponding Excel files. Since the tabular data cannot be used directly for model training, the obtained data were subjected to standardization preprocessing,

including missing value imputation, outlier removal, and normalization of continuous features, so as to eliminate the influence of measurement scales and improve training stability. Here, the MinMaxScaler utility of the scikit-learn library was used to perform Min-Max normalization on each feature column of the table, rescaling all numerical attributes to the range of 0 to 1 and thereby reducing the influence of differing scales and ranges among features.

$$X' = \frac{X - \mathrm{Min}(X)}{\mathrm{Max}(X) - \mathrm{Min}(X)} \quad (1)$$

Max(X) and Min(X) denote the maximum and minimum values in the column corresponding to feature X, respectively.

In addition, because the distribution of the patients' plantar pressure data cannot be observed directly, this study converted these data into heatmap images so that the subsequent model could effectively recognize such data. The code traverses each row of the table and extracts the corresponding raw plantar pressure data. The heatmap function of the Seaborn package was used to generate the visualization results, and various parameters (including the choice of colour map) can be customized by the user. Each heatmap image is uniquely identified by its corresponding class label. After generation, the system saves the heatmaps as image files to a designated folder according to the disease category and patient number. This integrated processing approach converts the raw pressure data into intuitive heatmap images that clearly present the distribution pattern of plantar pressure when the patient is standing.

$$H(M) = \mathrm{sns.heatmap}\{M, \mathrm{viridis}, \mathrm{xticklabels}, \mathrm{yticklabels}\} \quad (2)$$

M denotes the two-dimensional matrix of plantar pressure, viridis denotes the colour map used, and xticklabels and yticklabels denote the lateral sensor positions (along the plantar width direction) and the longitudinal sensor positions (along the plantar length direction), respectively. However, to better conform to the image format, the latter two parameters are set to False here.

Through the above preprocessing pipeline, missing values and outliers in the raw clinical data were effectively imputed and removed, the dimensional differences of continuous features were eliminated by Min-Max normalization, and the plantar pressure distribution data were uniformly transformed into standardized heatmap images. This integrated scheme not only significantly improves the quality of the dataset and the stability of model training, but also provides high-quality, directly usable input representations for the subsequent intelligent diagnostic model that fuses numerical features with image inputs, laying a solid data foundation for the accurate auxiliary diagnosis of musculoskeletal injuries.

2.1.3. Data Augmentation

In the field of clinical auxiliary diagnosis, data augmentation methods are mainly applied to small datasets or datasets in which the distribution of disease-category image samples is imbalanced, so as to increase the number of samples or to balance the dataset as much as possible. As deep learning models grow deeper and their parameter counts become increasingly large, data augmentation becomes particularly important for enabling small datasets to be trained effectively and for improving model recognition accuracy. Because the total number of samples in this study is relatively insufficient and the sample sizes of the disease categories are imbalanced, this paper adopts data augmentation methods such as mirroring, cropping, scaling, translation, and rotation. Without altering the core lesion content of the images, the spatial positions of pixels are changed through spatial geometric transformations, thereby effectively expanding the number of samples and preventing the model from overfitting during training. The musculoskeletal injury dataset used in this study contains clinical sample images under 4 different conditions, amounting to 702 images in total. Through the above data augmentation methods, the sample images of the imbalanced categories were all expanded to more than 450 images (the augmented dataset), which ultimately contains 1,903 images in total.

2.2. Model Construction

2.2.1. Improved CBAM (DCMSF-CBAM)

The attention mechanism is a mechanism that simulates human visual attention; it enhances the effectiveness of feature representation by focusing on important features and suppressing unnecessary ones. CBAM (Convolutional Block Attention Module) is a lightweight convolutional attention module[12]. Introducing the CBAM module into a convolutional neural network (CNN) can significantly improve model performance, while the additional parameters and computation introduced by the module are extremely small, so it can be conveniently embedded into most convolutional neural network architectures.

The information carried by different dimensions of a feature map differs markedly: the channel dimension emphasizes the semantic abstract representation of features, whereas the spatial dimension emphasizes the positional information of objects. CBAM injects attention maps along the two independent dimensions of channel and space of the feature map, and then multiplies the attention by the input feature map to perform adaptive feature refinement on the input feature map. The overall structure is shown in Fig. 2.

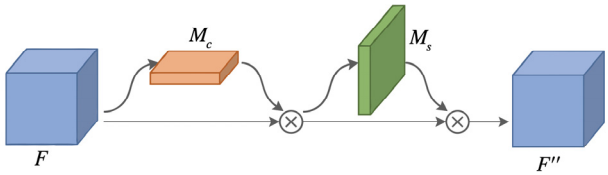


Fig 2. Structure of CBAM

F is given as the input intermediate feature map; the weighted result of the channel attention map F' is first obtained, and the spatial attention map M_s is then obtained through the spatial attention module, yielding the final weighted feature F''. The formulation is as follows:

$$\begin{cases} F' = M_c(F) \otimes F \\ F'' = M_s(F') \otimes F' \end{cases} \quad (3)$$

$M_c(F)$ is the output weight after channel attention, F is the intermediate feature map, $M_s(F')$ is the output weight after spatial attention, F' is the channel-refined feature map, and $\otimes$ denotes the element-wise multiplication operator.

Although CBAM enhances features through serially connected channel and spatial attention mechanisms, its channel attention fails to dynamically fuse cross-channel interaction information, and its spatial attention fails to capture multi-scale contextual information in parallel. Therefore, in order to achieve dynamic multi-path fusion of the two attention modules along both dimensions, this study improves CBAM and proposes a serial dynamic channel and spatial attention mechanism supplemented with a residual

connection, namely DCMSF-CBAM. The overall structure is shown in Fig. 3.

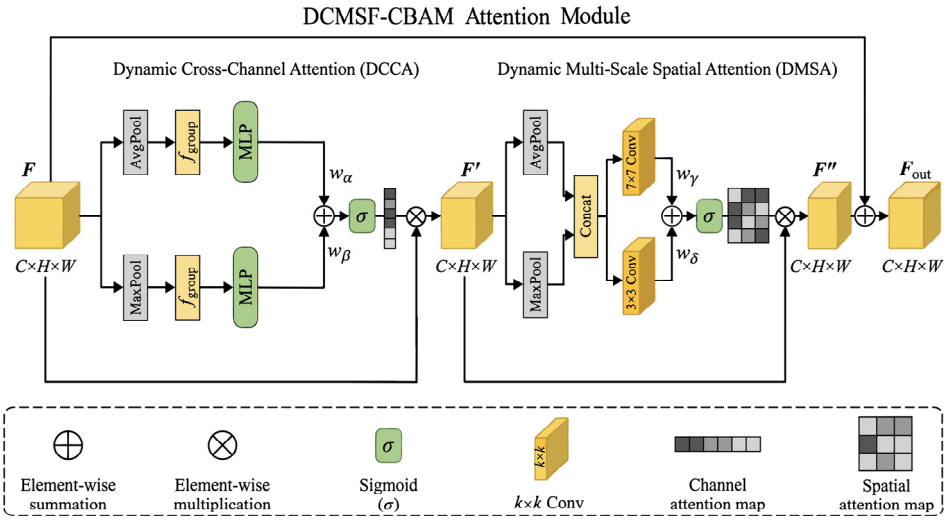


Fig 3. The DCMSF-CBAM attention module

$F \in \mathbb{R}^{C\times H\times W}$ is given as the intermediate feature map; the processing pipeline of this module is as follows:

1) Dynamic cross-channel attention computation: on the basis of aggregating spatial information through average pooling and max pooling, a cross-channel interaction module and learnable dynamic weight parameters are introduced to fuse the information of different pooling paths more flexibly and to generate the channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$. The formulation is as follows:

$$M_c(F) = \sigma\left(\frac{e^{\alpha}}{e^{\alpha}+e^{\beta}}\cdot \mathrm{MLP}(f_{group}(\mathrm{AvgPool}(F))) + \frac{e^{\beta}}{e^{\alpha}+e^{\beta}}\cdot \mathrm{MLP}(f_{group}(\mathrm{MaxPool}(F)))\right) \quad (4)$$

σ denotes the Sigmoid function, α and β denote learnable weight parameters, f_{group} denotes grouped convolution, and MLP is a shared multilayer perceptron whose hidden-layer dimension is $\mathbb{R}^{C/r\times 1\times 1}$, where r is the reduction ratio;

2) Dynamic multi-scale spatial attention computation: on the channel-refined feature map, average pooling and max pooling are performed along the channel axis, and the results are concatenated and fed into a parallel multi-scale convolution module (7×7 and 3×3 convolutions); learnable dynamic weight parameters are simultaneously introduced to adaptively and weightedly fuse the features extracted by the two paths at different scales, thereby generating the spatial attention map, where $F' = M_c(F) \otimes F$ and $M_s \in \mathbb{R}^{1\times H\times W}$. The formulation is as follows:

$$M_s(F') = \sigma\left(\frac{e^{\gamma}}{e^{\gamma}+e^{\delta}}\cdot f^{7\times 7}([\mathrm{AvgPool}(F');\mathrm{MaxPool}(F')]) + \frac{e^{\delta}}{e^{\gamma}+e^{\delta}}\cdot f^{3\times 3}([\mathrm{AvgPool}(F');\mathrm{MaxPool}(F')])\right) \quad (5)$$

$f^{7\times 7}$ and $f^{3\times 3}$ denote convolution operations with filter sizes of 7×7 and 3×3, respectively, AvgPool and MaxPool denote average pooling and max pooling, and γ and δ denote learnable weight parameters;

$F'' = M_s(F') \otimes F'$ is the spatially refined feature map; a residual connection is introduced to preserve the original information, yielding the final generated feature map $F_{out} \in \mathbb{R}^{C\times H\times W}$. The formulation is as follows:

$$F_{out} = F'' + F \quad (6)$$

The DCMSF-CBAM module enhances features through serially connected dynamic channel and spatial attention mechanisms. Its channel attention dynamically and weightedly fuses the grouped convolution (cross-channel interaction) path and the MLP path; its spatial attention, in turn, adaptively aggregates spatial information with dynamic weights from the contextual information captured in parallel by multi-scale convolutions. Supplemented by a residual connection, the module achieves dynamic multi-path fusion along both dimensions and performs flexible and fine-grained feature enhancement, significantly improving the model's feature representation capability for musculoskeletal injuries.

2.2.2. Interaction-Aware Gated Fusion Module (IAGF)

The core of a gating unit is to learn the importance weights of input features through one or more neural networks and to dynamically weight these features so as to adapt to different feature fusion scenarios. Specifically, this structure introduces the Sigmoid activation function to output a gating signal, strictly mapping the weight value of each feature channel into the interval from 0 to 1, thereby quantifying the degree to which the target feature is suppressed or enhanced. Subsequently, the gating weights thus produced are applied directly to the input features and combined with them through an element-wise multiplication operation, thereby completing the weighted regulation of different feature channels. Through this weighting mechanism, different feature channels receive different weights according to task requirements, which enables the model to adapt to different feature fusion situations[13]. In scenarios where visual features are important, the model may assign higher weights to visual information, whereas in regions where numerical features are critical, the model tends to strengthen the

numerical features. To better handle the different distributions of visual and numerical features and to effectively address the problems of incomplete fusion and loss of important information, this paper designs IAGF, whose overall structure is shown in Fig. 4.

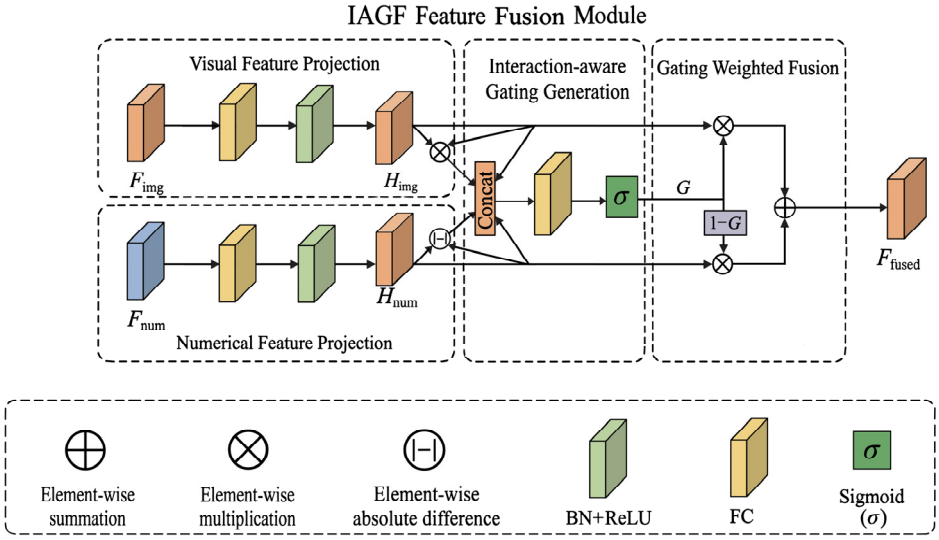


Fig 4. The IAGF feature fusion module

By learning dynamic gating coefficients, the IAGF fusion module can adaptively adjust the proportion of image features to numerical features according to the validity and reliability of different features, retaining more of the features of the current branch when attention is important and more of the features of the other branch when it is not, thereby enabling dynamic adjustment by the model. The visual features extracted from the input image through convolutional layers and residual blocks are denoted as $F_{img} \in \mathbb{R}^{B\times D_{img}}$, and the input numerical features are denoted as $F_{num} \in \mathbb{R}^{B\times N_{num}}$. The processing pipeline of this module is as follows:

1) Projection of features to a unified spatial dimension: the two types of features are mapped to the same spatial dimension through fully connected layers, where $D_{fusion} = 512$. The formulation is as follows:

$$H_{img} = \mathrm{ReLU}\left(\mathrm{BN}\left(W_{img}F_{img} + b_{img}\right)\right) \in \mathbb{R}^{B\times D_{fusion}} \quad (7)$$

$$H_{num} = \mathrm{ReLU}\left(\mathrm{BN}\left(W_{num}F_{num} + b_{num}\right)\right) \in \mathbb{R}^{B\times D_{fusion}} \quad (8)$$

$W_{img} \in \mathbb{R}^{D_{fusion}\times D_{img}}$ and $W_{num} \in \mathbb{R}^{D_{fusion}\times N'_{num}}$ are weight projection matrices, $N'_{num} = 256$ is the dimension of the numerical features after processing by the first two network layers;

2) Computation of the interaction-aware gating weights: two types of interaction features are first constructed. The element-wise product characterizes the synergistic consistency of the two features along the corresponding dimensions, whereas the element-wise absolute difference characterizes the discrepancy and potential conflict between the two features. The formulations are as follows:

$$P = H_{img} \otimes H_{num} \in \mathbb{R}^{B\times D_{fusion}} \quad (9)$$

$$\Delta = |H_{img} - H_{num}| \in \mathbb{R}^{B\times D_{fusion}} \quad (10)$$

The two projected features and the interaction features are then concatenated along the channel dimension, and a weight vector is generated through a gating network. The formulations are as follows:

$$U = \mathrm{Concat}\left(H_{img}, H_{num}, P, \Delta\right) \in \mathbb{R}^{B\times 4D_{fusion}} \quad (11)$$

$$G = \sigma\left(W_{gate}U + b_{gate}\right) \in \mathbb{R}^{B\times D_{fusion}} \quad (12)$$

σ is the Sigmoid function, which compresses the output to [0,1] and represents the degree of gating openness in each dimension, and $W_{gate} \in \mathbb{R}^{D_{fusion}\times 4D_{fusion}}$ is the weight of the gating fully connected layer;

3) Gated weighted fusion: the gating weights are used to perform weighted fusion of the two types of projected features. The formulation is as follows:

$$F_{fused} = G \otimes H_{img} + (1 - G) \otimes H_{num} \in \mathbb{R}^{B\times D_{fusion}} \quad (13)$$

$\otimes$ denotes element-wise multiplication; this formulation enables the model to adaptively adjust the contribution ratio of visual and numerical features according to the input content.

Through its learnable gating mechanism, the IAGF module achieves adaptive fusion of visual and numerical features. In the computation of the gating weights, it explicitly introduces two types of interaction representations, namely the element-wise product and the absolute difference, explicitly characterizing the synergistic consistency and the discrepancy/conflict between the two types of features, and it controls the two information flows through Sigmoid gating, so that the model can emphasize important features and suppress redundant or noisy information according to the specific input, thereby improving its ability to discriminate abnormal gait.

2.2.3. Dual-Branch Heterogeneous Network

The musculoskeletal injury model proposed in this paper adopts a dual-branch network architecture that performs extraction, encoding, and cross-modal fusion for gait image features and numerical features, respectively, so as to achieve accurate classification of disease types related to musculoskeletal injuries.

The image branch adopts Swin-T as the backbone network

and loads model weights pre-trained on the ImageNet dataset through transfer learning[14]. The input plantar pressure distribution map is first partitioned into non-overlapping patches by Patch Embedding (stride 4) and mapped into tokens. It then enters a four-level hierarchical Transformer (Stage 1-4), and DCMSF-CBAM is inserted after the output of each Stage: for Stages 1-3 it is placed before Patch Merging to alleviate information loss during downsampling, whereas for Stage 4 it is placed before Layer Norm to further strengthen the model's ability to focus on critical gait regions. Finally, compact high-level image features are obtained through Layer Norm and average pooling (Avg Pool). The numerical feature branch takes gait parameters such as gait speed, stride length, and joint range of motion as input and performs abstract encoding through two consecutive fully connected layers, realizing the nonlinear mapping and high-dimensional representation of low-dimensional numerical information.

In the feature fusion stage, the outputs of the two branches are adaptively and weightedly integrated across modalities through the IAGF module, fully exploiting the complementary relationship between image and numerical data. Subsequently, feature compression and deep interaction are completed through fully connected layers, and the prediction probabilities of each disease class are finally output by a Softmax classifier. The model framework is shown in Fig. 5.

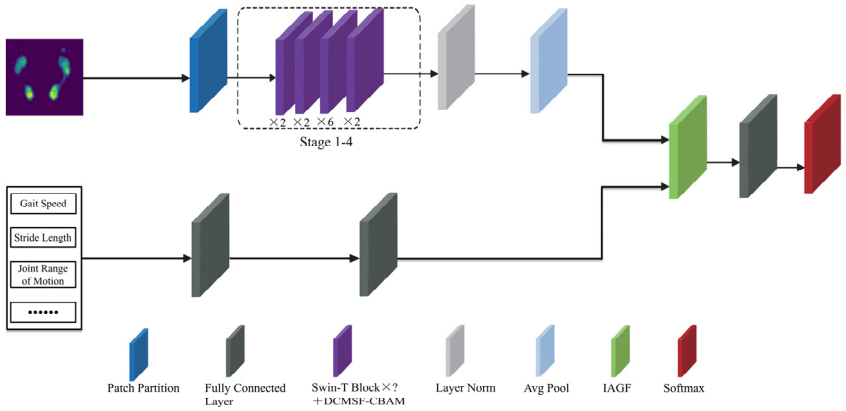


Fig 5. Overall architecture of the model

3. Experimental Results and Analysis

3.1. Experimental Environment

All experiments in this study were conducted under the Linux operating system. The server was equipped with an Intel(R) Xeon(R) Silver 4214 processor and 8 NVIDIA GeForce RTX 3080 graphics cards to provide computational acceleration. The training, validation, and testing procedures of the model were all implemented in the PyCharm 2024 integrated development environment, based on the Python 3.9 programming language and the PyTorch 1.13 deep learning framework.

3.2. Experimental Parameters and Configuration

In this study, the cross-entropy loss function (CrossEntropyLoss) was used to evaluate the model error, and the SGD algorithm with momentum was selected as the optimizer. In terms of parameters, the initial learning rate was initialized to 0.01, and the momentum and weight decay were configured as 0.9 and 1×10^{-4}, respectively. The prepared dataset was randomly divided into three parts, of which 75% was used for training, 15% for validation, and 10% for testing. The batch size and the total number of epochs for network training were set to 32 and 20, respectively.

3.3. Evaluation Metrics

In order to objectively and comprehensively quantify the performance of the proposed model in the musculoskeletal injury recognition task, this study not only adopts the global accuracy (Accuracy) to measure the overall prediction correctness of the model, but also introduces precision, recall, and F1-score based on the macro-average strategy.

The specific formulations adopted in this study are as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FN_i)} \tag{14}$$

$$\text{Precision}_{macro} = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i} \tag{15}$$

$$\text{Recall}_{macro} = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i} \tag{16}$$

$$\text{F1}_{macro} = \frac{1}{C}\sum_{i=1}^{C} \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{17}$$

C denotes the total number of disease categories, and TP_i, FP_i, and FN_i denote the true positives, false positives, and false negatives of the i-th class, respectively.

3.4. Model Comparison

In order to objectively evaluate the performance of the proposed model in the musculoskeletal injury recognition task, several mainstream deep learning models, namely MobileNetV2, Swin-T, and ConvNeXt V2 Tiny (ConvNeXt V2-T), were selected in the experiments and used to replace the visual branch of the proposed model that processes image features, while the numerical feature branch and the subsequent feature fusion structure were kept unchanged. The results are shown in Table 1. The proposed model achieves the best performance on all metrics, with Accuracy, Precision,

Recall, and F1-Score reaching 0.9476, 0.9456, 0.9557, and 0.9500, respectively, significantly outperforming the other three compared models. Compared with Swin-T, which ranks second, the above metrics are improved by 0.0628, 0.0650, 0.0675, and 0.0676, respectively. These results indicate that the proposed model can extract discriminative features more stably and accurately, and possesses stronger recognition capability and application potential in this task.

Table 1. Comparative experimental results

Model	Accuracy	Precision	Recall	F1-Score
MobileNetV2	0.8534	0.8469	0.8578	0.8472
Swin-T	0.8848	0.8806	0.8882	0.8824
ConvNeXt V2-T	0.8743	0.8671	0.8712	0.8685
Our Model	0.9476	0.9456	0.9557	0.9500

3.5. Model Ablation

In the musculoskeletal injury recognition task, in order to objectively verify the effectiveness of the proposed DCMSF-CBAM attention module and the IAGF feature fusion module, this study conducted systematic ablation experiments with Swin-T as the baseline model; the results are shown in Table 2. The Accuracy, Precision, Recall, and F1-Score of the baseline Swin-T model are 0.8429, 0.8448, 0.8402, and 0.8378, respectively. After the introduction of standard CBAM, the metrics of the Swin-T+CBAM model increase to 0.8691, 0.8678, 0.8637, and 0.8642, verifying the promoting effect of the attention mechanism on feature extraction. After replacing CBAM with the proposed DCMSF-CBAM, the performance of the Swin-T+DCMSF-CBAM model further improves to 0.9005, 0.8952, 0.9025, and 0.8981, indicating that the dynamic channel-spatial attention mechanism significantly enhances the feature representation capability. The Swin-T+IAGF model, which integrates IAGF alone, achieves 0.8848, 0.8806, 0.8882, and 0.8824, confirming the advantage of this gating mechanism in the adaptive fusion of visual and numerical features.

Ultimately, the proposed Our Model (Swin-T+DCMSF-CBAM+IAGF) achieves the best performance, with an Accuracy of 0.9476, a Precision of 0.9456, a Recall of 0.9557, and an F1-Score of 0.9500. This model embeds the DCMSF-CBAM module into the Swin-T backbone and adopts IAGF in place of the original simple concatenation in the feature fusion stage. The ablation results fully demonstrate that the two innovative modules act synergistically, significantly improving the recognition accuracy and generalization capability of the model for musculoskeletal injuries and providing reliable support for clinical auxiliary diagnosis.

Table 2. Ablation experimental results

Model	Accuracy	Precision	Recall	F1-Score
Swin-T	0.8429	0.8448	0.8402	0.8378
Swin-T+CBAM	0.8691	0.8678	0.8637	0.8642
Swin-T+DCMSF-CBAM	0.9005	0.8952	0.9025	0.8981
Swin-T+IAGF	0.8848	0.8806	0.8882	0.8824
Our Model	0.9476	0.9456	0.9557	0.9500

4. Conclusion and Outlook

To address the problems of conventional musculoskeletal injury diagnosis methods, such as strong subjectivity, dependence on expensive imaging modalities, and insufficient fusion of multimodal gait features, this paper proposes a novel dual-branch deep learning architecture for the auxiliary diagnosis of musculoskeletal injuries. The model independently extracts the spatial texture of plantar pressure heatmaps and the nonlinear features of quantified gait parameters through a dual-branch network, and innovatively designs the DCMSF-CBAM attention mechanism and the IAGF feature fusion module. On the one hand, DCMSF-CBAM achieves dynamic multi-path fusion along both the spatial and channel dimensions, significantly enhancing the network's fine-grained perception of critical lesion regions such as the heel and the arch; on the other hand, IAGF achieves adaptive weighted integration of visual representations and quantified numerical values by introducing a dynamic gating mechanism driven by interaction features, effectively overcoming the information redundancy or loss caused by conventional simple concatenation methods. Experimental results show that the Accuracy, Precision, Recall, and F1-Score of the model in the four-class task (healthy, knee osteoarthritis, anterior cruciate ligament injury, and meniscus injury) reach 0.9476, 0.9456, 0.9557, and 0.9500, respectively. All evaluation metrics are significantly superior to those of mainstream networks such as MobileNetV2, Swin-T, and ConvNeXt V2-T, and the ablation experiments fully verify the effectiveness and synergistic advantages of each core innovative component.

Although this study has achieved satisfactory experimental results, there remains room for further deepening in future work. First, although the current experimental data have been augmented, they all originate from a single medical institution and therefore suffer from certain regional and data-acquisition-device limitations. In the future, multiple centres will be jointly organized to build a larger-scale and more diverse multimodal gait database, so as to further validate and improve the cross-domain generalization capability of the model. Second, at the present stage the model mainly focuses on the fusion of static spatial features and global numerical features; in the future, the spatio-temporal graph convolutional network (STGCN) architecture may be introduced to deeply exploit the dynamic temporal evolution patterns within the gait cycle, and attempts may be made to fuse more physiological modality signals such as surface electromyography. Finally, lightweight model compression algorithms will be explored to promote the real-time deployment of this auxiliary diagnosis system on mobile terminals or smart wearable devices.

References

[1] Safiri, S., Kolahi, A. A., Smith, E., et al. (2020). Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the Global Burden of Disease Study 2017. Annals of the Rheumatic Diseases, 79(6), 819–828. https://doi.org/10.1136/annrheumdis-2019-216076.

[2] Ji, X. Y., Yu, S. Y., Sun, Y. Y., et al. (2026). Artificial intelligence and gait analysis based auxiliary diagnosis and treatment methods for orthopaedic diseases. Journal of Shandong University (Health Sciences), 64(2), 34–43, 65. (in Chinese).

[3] Teng, J. M., Ma, Y., Liu, D. W., et al. (2022). Markerless gait analysis system based on a deep learning fusion algorithm. Journal of Medical Biomechanics, 37(4), 733–740. (in Chinese)

[4] Bodepu, D., Junaid, M. I., Madarapu, S., et al. (2025). CRA-Net: Cross reverse attention network for classification of neuro-degenerative diseases based on gait analysis. Biomedical Signal Processing and Control, 108, 107886. https://doi.org/10.1016/j.bspc.2025.107886.

[5] Kaya, M., Karakuş, S., Tuncer, S. A. (2022). Detection of ataxia with hybrid convolutional neural network using static plantar pressure distribution model in patients with multiple sclerosis. Computer Methods and Programs in Biomedicine, 214, 106525. https://doi.org/10.1016/j.cmpb.2022.106525.

[6] Li, C. H., Yang, P., Feng, C. L., et al. (2025). Abnormal gait recognition of stroke patients based on deep learning fusion. Journal of Medical Biomechanics, 40(4), 955–962. (in Chinese)

[7] Atick, A. M. F., H. E, M. C., Bin, Z. M., et al. (2023). NDDNet: a deep learning model for predicting neurodegenerative diseases from gait pattern. Applied Intelligence, 53(17), 20034–20046. https://doi.org/10.1007/s10489-023-04547-1.

[8] Rios, R. Y. L., Tiznado, O. E. J., Acosta, R. R. E., et al. (2025). Toward Smarter Orthopedic Care: Classifying Plantar Footprints from RGB Images Using Vision Transformers and CNNs. Journal of Imaging, 11(11), 414. https://doi.org/10.3390/jimaging11110414.

[9] Kaya, E., Şirzai, H., Yavuzer, G., et al. (2025). Classification of knee osteoarthritis severity using markerless motion capture and long short-term memory fully convolutional network. Computers in Biology and Medicine, 195, 110647. https://doi.org/10.1016/j.compbiomed.2025.110647.

[10] Chakraborty, J., Nandy, A. (2020). Discrete wavelet transform based data representation in deep neural network for gait abnormality detection. Biomedical Signal Processing and Control, 62. https://doi.org/10.1016/j.bspc.2020.102123.

[11] Kwon, B. S., Han, S. H., Lee, C. M., et al. (2020). Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images. IEEE Access, 8, 120597–120603. https://doi.org/10.1109/ACCESS.2020.3005948.

[12] Woo, S., Park, J., Lee, J. Y., et al. (2018). Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-01234-2_1.

[13] Wang, Y., Chen, Y. Y., Liu, J. J., et al. (2025). Image dehazing algorithm based on Transformer and gated fusion mechanism. Computer Systems & Applications, 34(2), 1–10. (in Chinese)

[14] Raghu, M., Zhang, C., Kleinberg, J., et al. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1905.03120.

[15] Yu, B., Yin, H., Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875.