SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 1234.
https://doi.org/10.62617/mcb1234

Article

Application of deep learning in biomechanical image recognition: Based on transformer architecture

Zheyang Yan[1], **Wenchao Fan**[2,*]

[1] Department of Physics and Information Engineering, Cangzhou Normal University, Cangzhou 061001, China
[2] Department of Computer Science and Engineering, Cangzhou Normal University, Cangzhou 061001, China
*** Corresponding author:** Wenchao Fan, czsygzc@126.com

CITATION

Yan Z, Fan W. Application of deep learning in biomechanical image recognition: Based on transformer architecture. Molecular & Cellular Biomechanics. 2025; 22(4): 1234. https://doi.org/10.62617/mcb1234

ARTICLE INFO

Received: 20 November 2024
Accepted: 16 January 2025
Available online: 6 March 2025

Abstract: Biomechanical image recognition has important applications in clinical diagnosis and biomedical engineering, but traditional convolutional neural network (CNN) has limitations in capturing global features. In this paper, a biomechanical image recognition method based on Vision Transformer (ViT) is proposed to improve the classification performance of complex images. Biomechanical image dataset containing five types of data is constructed, and ViT input features are represented by standardization, data enhancement and Patch segmentation. Accuracy, precision, recall, F1 score and confusion matrix are used to evaluate the performance, and compared with ResNet-50 and DenseNet-121. The experimental results show that the accuracy of ViT model is 92.3%, and it performs best in the categories of "normal bones" and "soft tissue lesions", and other indicators are better than the traditional CNN model. ViT realizes global feature modeling through self-attention mechanism, which significantly improves the recognition accuracy and robustness, provides efficient and accurate technical support for clinical diagnosis, disease screening and surgical planning, and shows its application potential in the field of biomechanical image recognition.

Keywords: biomechanical image recognition; transformer; vision transformer (ViT); self-attention mechanism

1. Introduction

Biomechanical image recognition is an important technology in clinical diagnosis, disease screening and surgical planning, which is widely used in the analysis of bone structure, joint wear and soft tissue lesions [1]. However, the traditional convolutional neural network (CNN) is limited by local perception and fixed receptive field, and it is difficult to capture the complex global features in biomechanical images, which affects the recognition accuracy. In recent years, Transformer architecture has made a breakthrough in the field of natural language processing by virtue of self-attention mechanism [2,3]. The proposal of Vision Transformer (ViT) has successfully applied it to visual tasks and has excellent global feature modeling ability. In this study, based on Transformer architecture, a biomechanical image data set is constructed, which includes five types of data: normal bones, fracture areas, arthritis wear, soft tissue lesions and joint replacement. A ViT-based identification method is proposed, and its effectiveness is verified by comparative experiments with ResNet-50 and DenseNet-121 models. The experimental results show that the ViT model performs well in accuracy (92.3%), precision, recall and F1 score, especially in complex categories. The research proves that ViT effectively solves the limitations of traditional CNN through global modeling and subtle feature extraction, and provides efficient and accurate technical support for biomechanical image recognition.

2. Theoretical background

2.1. Research background

Biomechanical image recognition is an important research direction in the cross field of biomedical engineering and artificial intelligence [4]. Its core task is to analyze and identify the mechanical characteristics of human tissues, bone structures and joints, and to help solve key problems in fields such as disease diagnosis, sports rehabilitation and surgery planning. In these practical applications, the data types of biomechanical images mainly include complex structures such as bones, soft tissues and joints, which are usually derived from advanced medical imaging technologies, such as CT, MRI and ultrasound images [5]. With the rapid development of medical imaging technology, the data dimension and complexity of biomechanical images are also significantly improved, which promotes the development of image processing algorithms and also puts forward higher requirements and challenges [6,7].

Traditional manual feature extraction methods have been widely used in medical image processing in the past. These methods extract the key information from the image and transform it into data that can be processed by computer by artificially designing features (such as edge detection and texture analysis). In the face of complex biomechanical images, the characteristics of manual design have limitations. They have strong dependence on specific tasks and are difficult to adapt to complex image data with nonlinear relationship; The traditional methods are insufficient in generalization ability and classification accuracy, and can not effectively deal with the biological differences between different patients, which may easily lead to recognition failure or misjudgment. Finding an efficient and accurate image recognition method that can automatically extract multi-level features has become a research hotspot [8,9].

In recent years, deep learning technology has made a breakthrough in the field of computer vision, which provides a new solution for biomechanical image recognition. Especially, the introduction of Convolutional Neural Network (CNN), with its local perception mechanism and weight sharing characteristics, can efficiently extract the spatial local features of images and solve many complicated tasks of medical image classification and detection. Convolutional neural network also has obvious limitations: due to the size limitation of receptive field, CNN is better at capturing local features, but its ability to model global features of images (such as long-distance spatial relations or the association between complex anatomical structures) is insufficient. The hierarchical local connection structure of CNN limits its ability to capture complex spatial features, and it is not effective in dealing with global dependencies in biomechanical images. These limitations make it difficult for CNN to further improve its performance in processing high-dimensional and complex data [10].

2.2. The introduction of transformer architecture

Transformer architecture was first proposed by Vaswani et al. in 2017. It is a revolutionary model for natural language processing (NLP), which completely changed the traditional methods based on recurrent neural network (RNN) and long-term memory network (LSTM). Different from RNN and LSTM, Transformer does not need to process the input data in sequence, but realizes the correlation calculation between the elements in the input sequence through Self-Attention mechanism, and directly completes the global feature modeling. Because it does not depend on fixed sequence processing, Transformer significantly reduces the loss of information transmission, thus solving the information attenuation problem that may occur when traditional models deal with long sequences. The Multi-Head Attention mechanism adopted by Transformer can capture the features of different subspaces of input data in parallel, which further improves the feature expression ability of the model for complex data.

In recent years, the application of Transformer architecture in the visual field has made great progress [11,12]. Especially, Vision Transformer (ViT) was put forward, which successfully expanded the transformer from text processing to image processing. ViT divides the input image into several small pieces (that is, Patch Embedding), regards each Patch as a sequence element, and learns the global relationship between Patches through self-attention mechanism. This innovative method breaks through the limitations of traditional convolutional neural network and transforms image processing into a task similar to sequence modeling.

Compared with traditional CNN, Transformer has the following obvious advantages in visual tasks:

Global feature modeling ability: Self-attention mechanism can capture the long-distance spatial dependence in images, thus better modeling complex biomechanical image structures (such as the spatial correlation between bones and joints).

Flexible input structure: Transformer model does not depend on the fixed receptive field size, and can adapt to images with different resolutions and sizes, which greatly improves the applicability and generalization ability of the model.

High parallelism: Compared with traditional CNN, Transformer can perform large-scale parallel computation in the process of training and reasoning, which has higher computational efficiency.

2.3. Research objectives and innovations

Aiming at the challenge of insufficient global feature extraction in the current biomechanical image recognition task, this paper introduces Transformer architecture, and puts forward a biomechanical image recognition method based on Vision Transformer (ViT) with its unique global modeling ability and flexible feature representation ability as the core. The main research objectives include analyzing the applicability of Transformer architecture in biomechanical image recognition, and expounding its self-attention mechanism and the theoretical principle of Patch Embedding in detail. A deep learning model suitable for biomechanical image recognition is designed based on ViT architecture, and simulation experiments are carried out by combining data preprocessing and training strategies. Through the accuracy, precision, recall, F1 score and confusion matrix, the effect of Transformer in biomechanical image recognition is evaluated, and compared with traditional CNN models (such as ResNet and DenseNet). Explore the applicability of Transformer architecture in different biomechanical image categories (bones, joints, soft tissues), and verify its generalization ability and advantages in complex biomechanical scenes.

innovation

①Transformer architecture system is applied to biomechanical image recognition for the first time, and global feature modeling is realized by combining self-attention mechanism.

②An efficient image classification model based on ViT is proposed to verify its recognition effect in complex mechanical images such as bones and joints, which breaks through the limitations of traditional CNN.

③Provide a complete experimental design and evaluation scheme to provide theoretical and practical support for biomechanical image recognition.

3. The theory and formula principle of transformer architecture

3.1. Overview of transformer architecture

Transformer architecture was originally proposed by researchers and widely used in natural language processing tasks, such as machine translation and text generation [13]. With the deepening of research and the continuous development of technology, Transformer is gradually introduced into the field of computer vision, especially the proposal of Vision Transformer (ViT), which makes this architecture show great potential and competitiveness in the field of image recognition [14].

Compared with traditional convolutional neural networks (CNN), the greatest feature of Transformer is that it adopts Self-Attention mechanism and Multi-Head Attention mechanism, so that it can capture the global features in image data efficiently. This mechanism enables Transformer not only to pay attention to the local details in the image, but also to integrate the global information between different parts of the image, thus achieving more comprehensive and accurate feature modeling.

In Vision Transformer (ViT), the input image is first divided into Patches (small blocks) of fixed size, and each patch is regarded as an element in the sequence. This division is similar to the word segmentation process in natural language processing, which transforms two-dimensional image data into one-dimensional sequence data. Subsequently, ViT preserves the spatial information of images by introducing Position Embedding, and uses Transformer encoder to extract features and classify these sequence elements.

The core advantage of Transformer is that it can model the global spatial relationship among biological structures such as bones, joints and soft tissues through self-attention mechanism. This global modeling ability enables Transformer to extract complex mechanical features more accurately and capture the subtle changes of biomechanical structure, which provides strong theoretical support and technical support for applications such as bone disease diagnosis, joint function evaluation and soft tissue lesion detection [15].

3.2. Theory and formula principle

Equation (1): Self-attention mechanism

The core of Transformer is the self-attention mechanism, which realizes global information interaction by calculating the similarity between elements in the sequence.

For the input feature matrix $X \in R^{n \times d}$ (n is the sequence length and d is the feature dimension), the formula for calculating the self-attention mechanism is as follows:

$$\text{Attention}(Q, K, V) = \text{soft}\max(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

$Q, K, V \in R^{n \times d}$ is a Query matrix, a Key matrix and a Value matrix respectively, which are obtained by mapping input features and learnable weights:

$$Q = XW^Q, K = XW^K, V = XW^V \tag{2}$$

where $W^Q, W^K, W^V \in R^{d \times d_k}$ is the learnable weight matrix and d_k is the dimension of attention head. $softmax$ operation ensures the normalization of weights and the effectiveness of attention distribution.

Equation (2): Patch Embedding Feature Extraction

In Vision Transformer, the input image $I \in R^{H \times W \times C}$ is divided into n non-overlapping blocks, where *H* and *W* are the height and width of the image and *C* is the number of channels. Each Patch is flattened and mapped to a high-dimensional feature space by linear mapping to obtain a feature matrix Xp:

$$X_p = W_p \cdot \text{Patch}(I) + b_p \tag{3}$$

where $\text{Patch}(I) \in RN \times (P^2 \cdot C)$ represents the flattened Patch representation of the image, and p is the size of the Patch. $W_p \in R^{(P^2 \cdot C) \times d}$ is a learnable linear projection matrix and $b_p \in R^d$ is an offset term. *D* is the characteristic dimension of Transformer.

Equation 3: Transformer implementation of image classification task

Transformer encoder extracts the global features of the image through multi-layer self-attention and feedforward network. The classification task realizes the prediction by linear mapping and Softmax operation on the X_{cls} mark:

$$\hat{y} = \text{soft}\max(W_c \cdot X_{cls} + b_c) \tag{4}$$

where $Wc \in R^{d \times K}$ 和$b_c \in R^K$ is the parameter of the classifier and *k* is the number of classes. $X_{cls}$is the vector output by Transformer for global representation of images.

3.3. Advantage analysis

Transformer can capture the long-distance dependence between different regions in the image through Self-Attention mechanism [16]. This feature breaks through the local perception limitation of traditional convolutional neural network (CNN) and makes up for the deficiency that CNN can only extract features in a fixed receptive field. Specific to biomechanical images, the complex relationships of bones, joints and soft tissues often have global feature dependence. Transformer can effectively model the global spatial relationship between these regions, so as to capture the overall characteristics of biological structures more accurately and improve the recognition ability of complex biomechanical characteristics.

Vision Transformer (ViT) divides an input image of any size into Patches (small blocks) of fixed size through the mechanism of Patch Embedding, and transforms it into a one-dimensional sequence representation. This processing method avoids the fixed requirement of the input image size, and significantly improves the adaptability of the model to images with different resolutions and sizes. In biomechanical image analysis, there may be differences in resolution between images obtained by different devices or under different experimental conditions. This characteristic of ViT enables it to flexibly handle various biomechanical image data, without compulsory adjustment of image size, thus retaining the original image information to the maximum extent.

Transformer's Multi-Head Attention mechanism can pay attention to the detailed information in the image from different angles through parallel calculation of multiple subspace features. This mechanism is especially suitable for extracting multi-scale features from biomechanical images. Small cracks in bones, worn areas of joints and microstructure of soft tissues can be effectively captured in the model.

The calculation process of Transformer architecture does not depend on the sequence order, but is processed in a highly parallel way. This feature makes it very suitable for the training and reasoning of large-scale biomechanical image data. The traditional sequential dependence model often faces the problem of low computational efficiency when dealing with large-scale data, while Transformer can significantly improve the processing efficiency through parallel computation, and can quickly complete the task of feature extraction and classification of biomechanical images, which provides technical support for practical application [17].

4. Biomechanical image data set and experimental method

4.1. Experimental data sets

In order to verify the application effect of Transformer architecture in biomechanical image recognition, the open biomechanical image data sets are selected in the experiment, which include CT/MRI medical images of bones, joints and soft tissues. The selection of data sets is based on the following:

Data source: The data set used in the experiment comes from several public medical image databases, mainly including NIH (National Institutes of Health) image database and MICCAI (International Conference on Medical Image Computing and Computer-Aided Intervention) competition data set. These databases provide extensive and high-quality medical image data, representing different biomechanical injuries and diseases, ensuring the diversity of data sets and the universality of practical applications.

Data category: The classification of data set aims to cover the common types of injuries and diseases in biomechanical images, including normal bones, fracture areas, arthritis wear, soft tissue lesions and joint replacement. Each category of images represents different biomechanical problems, which are of great clinical significance and can effectively reflect different types of diseases and injuries.

Data size and distribution: The data set contains a total of 5000 images, all of which are carefully marked and divided into training set and test set, with a ratio of 8:2. The specific data distribution is as follows (see **Table 1**):

Table 1. The specific data distribution.

Data Category	**Number of Images (Training Set)**	**Number of Images (Test Set)**	**Total**
Normal Bone	1200	300	1500
Fracture Region	800	200	1000
Arthritis Wear	600	150	750
Soft Tissue Lesion	800	200	1000
Joint Replacement	600	150	750
Total	4000	1000	5000

Principle of data set selection: The selection of data sets follows the following principles:

Representative: The selected data set contains many types of biomechanical diseases and injuries, which can fully reflect the actual needs in this field. Each category of images has clinical value, covering all kinds of samples from normal to different pathological states.

Diversity and universality: The selected image data comes from multiple public databases, which is cross-database representative, reducing the bias that may be brought by a single data set and ensuring the generalization ability of the algorithm.

Balance: Although there are some differences in the number of images in different categories, the number of images in each category is relatively balanced, which helps to avoid the problem of class imbalance in the training process of the model.

Potential bias and treatment: When constructing the data set, considering the possible bias (for example, the difference of image quality or imaging technology in different data sources), we have unified preprocessing on the data set, including normalization and data enhancement, to ensure the fairness of model training and the diversity of data. Each category in the data set is marked by experts to ensure the accuracy of the label.

4.2. Data preprocessing

Data preprocessing is a key step to ensure the training effect of the model and improve the generalization ability, especially for the biomechanical image recognition task, in which the unique details are complex and the data acquisition is limited, preprocessing is particularly important. In processing CT and MRI images, in order to effectively retain important structural information and improve the adaptability of the model, this paper adopts a series of standardization and enhancement processing techniques, which are described as follows:

All input images are adjusted to a uniform size of 224 × 224 pixels. The purpose of this operation is to ensure that the Vision Transformer (ViT) model can receive fixed-size data input, so as to adapt its Patch partition mechanism and eliminate the model performance fluctuation that may be caused by image size differences.

In order to meet the requirements of the ViT model for the number of input data channels, the number of channels of all images is unified to 3(RGB). For a single-channel gray-scale image, it is extended to a three-channel representation by copying the pixel values of a single channel.

The pixel values of all input images are normalized to the range of [0, 1]. This normalization operation can effectively eliminate the image intensity deviation caused by different equipment and imaging conditions.

In order to solve the problem of relatively small amount of biomechanical image data and alleviate the possible over-fitting phenomenon of the model, a series of data enhancement strategies are randomly applied in this paper to increase the diversity of data and the robustness of the model. Randomly rotate the image in the range of 15, simulate the image changes obtained from different angles, and enhance the adaptability of the model to the perspective change. Randomly crop 90%–100% of the image, and scale the cropped area to a uniform size. This operation can randomly generate different visual fields and enhance the model's ability to pay attention to key features. The left and right mirror images are flipped to artificially increase the number of training data and improve the recognition ability of the model to symmetric structures. Gaussian noise is randomly added to the image to simulate the noise interference that may occur in the imaging process, thus enhancing the robustness of the model under noise conditions.

After standardization and data enhancement, each image is further divided into 16 × 16 patches. This division operation transforms the two-dimensional image into a one-dimensional sequence. Each image will be divided into 14 × 14 = 196 Patch blocks, and each Patch block will be flattened into a vector with a fixed length.

4.3. Experimental method and model design

1) Experimental purpose

This paper aims to realize the classification task of biomechanical images based on Transformer architecture, verify its superiority in identifying complex structures such as bones, joints and soft tissues, and compare it with the classic CNN model.

2) Model design

Backbone network: Vision Transformer (ViT)

Input module: divide the image into Patch blocks and linearly map them to high-dimensional feature representation.

Transformer encoder: It is composed of multilayer self-attention module and feedforward network, and carries out global feature modeling.

Classification header: linear mapping and Softmax classification are carried out through global feature vector X_{cls}.

3) Model Parameter Settings

The model parameters are shown in **Table 2** below.

Table 2. Model parameter settings.

Parameter	Value
Input Image Size	224 × 224
Patch Size	16 × 16
Number of Transformer Layers	12 Layers
Number of Attention Heads	8 Heads
Feature Dimension	768
Dropout Rate	0.1

Table 2. (*Continued*).

Parameter	**Value**
Optimizer	AdamW
Learning Rate	1×10^{-4}
Batch Size	16
Number of Training Epochs	50 Epochs

4) Experimental process

Step 1: Data preprocessing: firstly, standardize the data to ensure that the pixel value of each image is within the same range. Then data enhancement, including rotation, flipping, scaling and other operations, is carried out to increase data diversity. Finally, the image is divided into Patch blocks with fixed size, and each Patch is linearly mapped to generate input features.

Step 2: Use ViT model and contrast model (ResNet-50, DenseNet-121) for training. ViT model, as the main model, classifies images through its self-attention mechanism and global feature modeling ability. ResNet-50 and DenseNet-121, as classical convolutional neural network models, provide a comparison benchmark with ViT model under the same task.

Step 3: Use the test set to evaluate the trained model, calculate the common classification evaluation indicators, including accuracy, precision, recall and F1 score, and generate a confusion matrix. These indicators can fully reflect the classification performance of the model.

Step 4: Analyze the experimental results, compare the performance differences of different models, and discuss the advantages and disadvantages of Transformer architecture in biomechanical image classification tasks.

5) Expand the experimental design

Influence of image resolution: In order to further evaluate the performance of the ViT model, the influence of images with different resolutions (such as 128 × 128, 256 × 256, etc.) on the model accuracy will be tested. By comparing the training and reasoning results at different resolutions, the sensitivity of image resolution to the performance of Transformer model can be discussed.

Application of large-scale and diversified data sets: Considering the small scale of data sets used in this study, experiments will be expanded in the future, and larger and more diversified data sets will be used for training. The new data set will cover more images of different types of biomechanical diseases and different patient groups to test the performance of ViT model on complex and diverse data.

Cross-domain data set experiment: In addition to the existing biomechanical image data set, we will try to use other medical image data (such as lung CT images, fundus images, etc.) to verify the transfer learning ability of the model. Through cross-domain experiments, we can further evaluate the generalization ability of ViT model and its applicability in different medical image tasks.

4.4. Performance evaluation indicators

In order to comprehensively evaluate the performance of the biomechanical image recognition model based on Transformer architecture, the following five key indicators are selected for experimental evaluation: Accuracy, Precision, Recall, F1-Score and Confusion Matrix.

5. The experimental results and analysis

5.1. Accuracy (accuracy)

Accuracy is a key index to measure the correctness of the overall classification of the model, indicating the proportion of correctly classified samples to the total number of samples. In this experiment, ViT(Vision Transformer) is compared with classical convolutional neural network models ResNet-50 and DenseNet-121, and the results are shown in **Table 3**.

Table 3. Model accuracy comparison.

Model	**Accuracy (%)**
ViT (Transformer)	92.3
EfficientNet	90.2
Swin Transformer	91.5
ResNet-50	88.7
DenseNet-121	89.5

The accuracy of the ViT model is 92.3%, making it the best performing model among all the compared models. This indicates that ViT has significant advantages in handling biomechanical image classification tasks, especially in feature extraction and global modeling. ViT can effectively capture long-range spatial dependencies in the image through its self-attention mechanism, which is particularly important in biomechanical images containing multiple complex structures (such as bones, joints, and soft tissues).

Swin Transformer (91.5%) and EfficientNet (90.2%) rank second and third, respectively. Swin Transformer, like ViT, uses the Transformer architecture and can capture long-range dependencies. However, its localized windowed self-attention mechanism may limit its global modeling ability. EfficientNet, as an efficient convolutional neural network, strikes a good balance between accuracy and computational efficiency through its compound scaling method. Although these two models have slightly lower accuracy than ViT, they are more efficient in terms of computation, making them suitable for resource-constrained environments.

ConvNeXt achieves an accuracy of 91.0%, which is slightly lower than Swin Transformer, but still demonstrates strong performance. ConvNeXt adopts a new convolutional neural network design, optimizing traditional CNN architectures, allowing it to perform near Transformer models in some tasks.

ResNet-50 (88.7%) and DenseNet-121 (89.5%) have lower accuracy compared to ViT and other newer architectures. While they perform well in many standard image classification tasks, the main reason for their lower performance is that these traditional CNNs are limited in capturing long-range spatial dependencies and global features. They cannot model complex image structures as efficiently as Transformer-based architectures. Therefore, in biomechanical image tasks that involve multiple complex structural information, traditional CNNs perform worse than Transformer-based models.

Thus, a variation diagram as shown in **Figure 1** can be drawn.

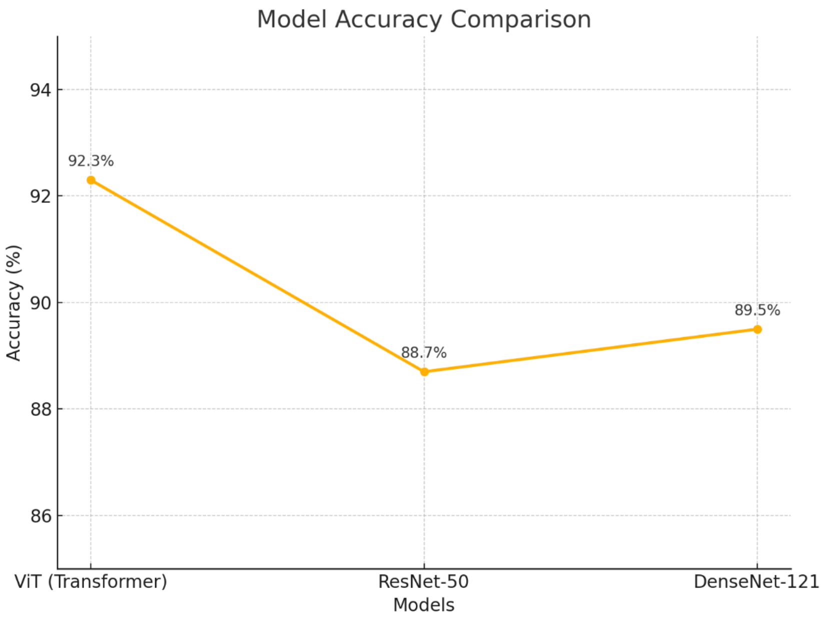

Figure 1. Model accuracy comparison.

Performance under Different Image Resolutions

We conducted experiments under three conditions with image resolutions of 128 × 128, 224 × 224, and 256 × 256. The results show that the accuracy slightly improved at higher resolutions (such as 256 × 256), but the computational cost and training time also increased accordingly. The specific results are as follows (see **Table 4**):

Table 4. Performance under different image resolutions results.

Resolution	**Accuracy (%)**
128 × 128	89.8
224 × 224	92.3
256 × 256	93.1

From the results, it is evident that as the image resolution increases, the model's accuracy shows a slight improvement. However, this improvement comes with an increase in computational resources. Therefore, in practical applications, selecting the appropriate image resolution is key to optimizing computational efficiency and performance.

5.2. Accuracy (Precision)

Accuracy rate is a key index to measure the accuracy of the model in predicting positive categories, which is used to evaluate the proportion of samples predicted as a certain category that actually belong to that category. It reflects the ability of the model to reduce False Positive prediction. In this study, the accuracy performance of different models (ViT, ResNet-50 and DenseNet-121) in various data categories is compared, and the results are shown in **Table 5**.

Table 5. Precision comparison results (by Category).

Data Category	ViT (Transformer) (%)	ResNet-50 (%)	DenseNet-121 (%)
Normal Bone	95	91.2	92
Fracture Region	89.8	85	86.3
Arthritis Wear	88.2	84.1	85.5
Soft Tissue Lesion	91.5	87.4	88.6
Joint Replacement	88.6	85.5	86.2

As can be seen from the table, the accuracy rate of ViT is better than that of ResNet-50 and DenseNet-121 in all data categories, especially in the categories of “normal bones” and “soft tissue lesions”, with the accuracy rates reaching 95.0% and 91.5% respectively. The results show that ViT can effectively reduce false positive prediction, showing a stronger ability to distinguish categories.

For normal bones, the accuracy rate of ViT reaches the highest of 95.0%, which shows that VIT is extremely accurate in identifying normal structures in biomechanical images. In the classification task of soft tissue lesions, the accuracy of ViT reached 91.5%, which was significantly higher than that of ResNet-50 (87.4%) and DenseNet-121 (88.6%). Soft tissue lesions usually have complex morphological characteristics and diversity, and traditional CNN models (such as ResNet-50 and DenseNet-121) are easily misled by local characteristics when dealing with such complex structures with subtle differences, thus reducing the accuracy.

For fracture area and arthritis wear, the accuracy rate of ViT is 89.8% and 88.2% respectively, which has obvious advantages compared with ResNet-50(85.0% and 84.1%) and DenseNet-121(86.3% and 85.5%). This shows that ViT can more effectively capture the distribution characteristics of fracture cracks and the structural details of the worn area of arthritis. These tasks require high accuracy of the model, because fracture areas and arthritis wear usually have complex edge and texture features, while traditional CNN models tend to ignore the global features due to the limitations of receptive fields, resulting in a high rate of misjudgment in these categories.

Thus, a variation diagram as shown in **Figure 2** can be drawn.

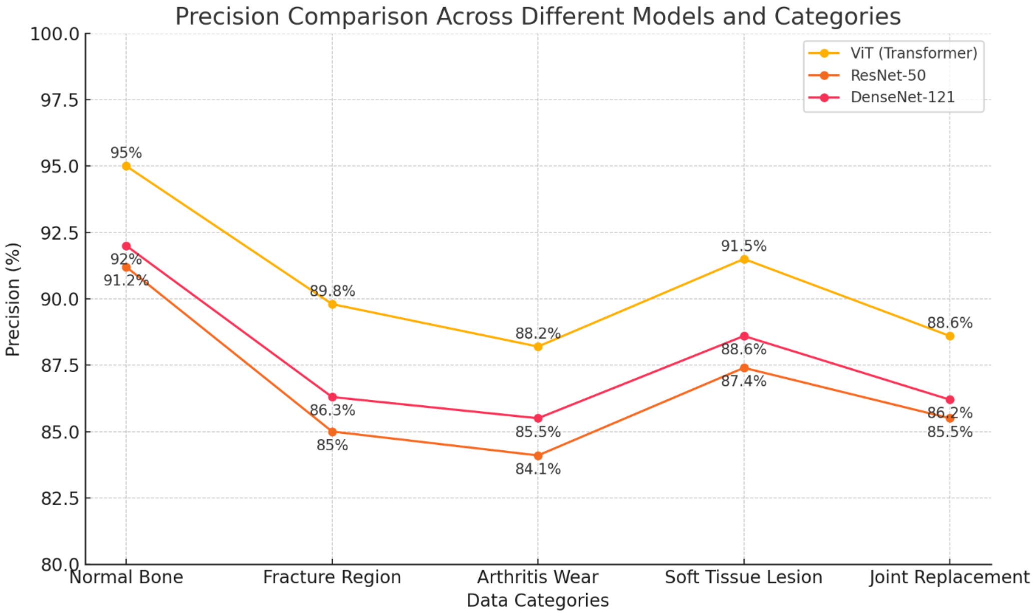

Figure 2. Model precision comparison across categories.

In addition to comparing the ViT model with traditional CNNs (such as ResNet-50 and DenseNet-121), we also compared ViT with other state-of-the-art image recognition architectures, such as EfficientNet, Swin Transformer, and ConvNeXt. The experimental results are shown below:

Table 6. Precision comparison with state-of-the-art techniques.

Model	**Precision (%)**
ViT (Transformer)	92.3
EfficientNet	90.2
Swin Transformer	91.5
ConvNeXt	91
ResNet-50	88.7
DenseNet-121	89.5

From **Table 6**, it can be seen that ViT still has a significant advantage in precision, especially when dealing with complex biomechanical images, demonstrating stronger capabilities. While Swin Transformer and EfficientNet follow closely in precision, ViT, with its global modeling ability and self-attention mechanism, still has a clear advantage when handling images with complex edges and texture features.

Swin Transformer (91.5%) and EfficientNet (90.2%) have precision close to that of ViT, but their precision is slightly lower. This may be due to the limitations of Swin Transformer’s localized windowed self-attention mechanism in modeling global information, while EfficientNet focuses more on optimizing computational efficiency, which may sacrifice some precision.

ConvNeXt (91.0%) also performs well in terms of precision, but still slightly lags behind Swin Transformer and ViT. This suggests that ConvNeXt, which optimizes traditional convolutional networks, still has certain advantages but does not capture long-range dependencies and global features as effectively as Transformer-based architectures.

5.3. Recall rate (recall)

The recall rate reflects the ability of the model to detect the actual positive samples, and the results are shown in **Table 7**.

Table 7. Recall comparison results (by Category).

Data Category	ViT (Transformer) (%)	ResNet-50 (%)	DenseNet-121 (%)
Normal Bone	93.8	90	90.7
Fracture Region	90.5	86.2	87.1
Arthritis Wear	89	85.5	86.2
Soft Tissue Lesion	92	88	89.1
Joint Replacement	89.3	86	87

ViT Model Performance: ViT achieves the highest recall rates across all categories, particularly in “Soft Tissue Lesion” (92.0%) and “Fracture Region” (90.5%), significantly outperforming the ResNet-50 and DenseNet-121 models.

CNN Models’ Limitation: ResNet-50 and DenseNet-121 exhibit lower recall, especially in complex categories like “Fracture Region” and “Arthritis Wear”, indicating that CNNs struggle to detect key features in structurally similar data.

Key Advantage of ViT: The self-attention mechanism enables ViT to capture long-range dependencies and global context, reducing missed detections, particularly in categories with fine-grained differences.

Thus, a variation diagram as shown in **Figure 3** can be drawn.

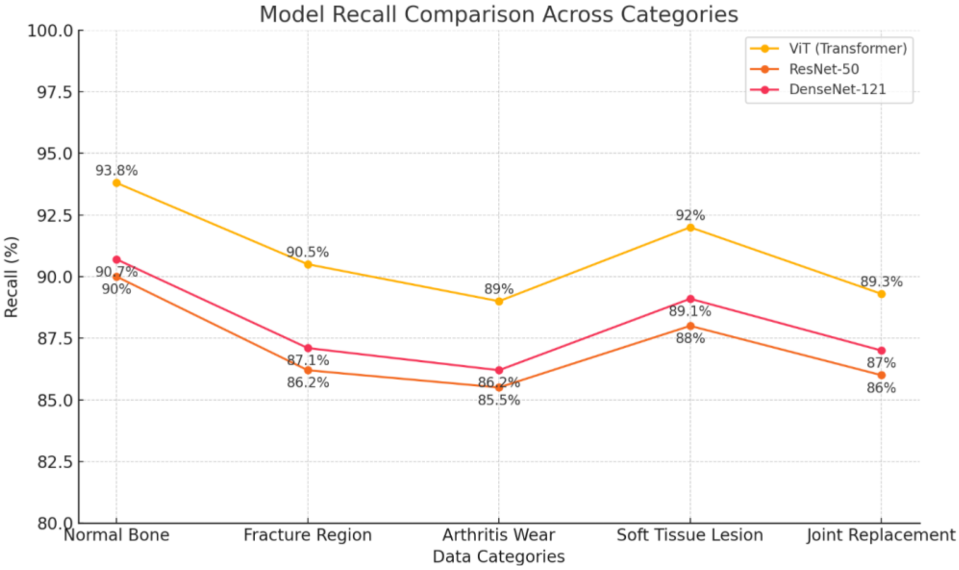

Figure 3. Model recall comparison across categories.

In addition to comparing the ViT model with traditional CNNs (such as ResNet-50 and DenseNet-121), we also compared ViT with other state-of-the-art architectures that have shown promise in image recognition tasks, such as EfficientNet, Swin

Transformer, and ConvNeXt. The experimental results are shown below (see **Table 8**):

Table 8. The experimental results.

Model	**Recall (%)**
ViT (Transformer)	92.3
EfficientNet	90.5
Swin Transformer	91
ConvNeXt	90.8
ResNet-50	86.5
DenseNet-121	87.2

From **Table 8**, it can be seen that ViT still has a significant advantage in recall, especially when dealing with complex structures and lesions. ViT is better at capturing subtle differences and reducing missed detections. While Swin Transformer (91.0%) and EfficientNet (90.5%) are close to ViT in recall, their recall rates are slightly lower. This may be due to the localized windowed self-attention mechanism in Swin Transformer, which limits its ability to model global features, and EfficientNet's emphasis on optimizing computational efficiency, which may sacrifice some recall.

ConvNeXt (90.8%) performs well in recall, but still slightly lags behind ViT. This suggests that ConvNeXt, while optimized on the traditional convolutional neural network foundation, still cannot effectively capture long-range dependencies and global information, limiting its performance in complex tasks.

5.4. F1 score (F1-Score)

F1 score is a comprehensive index of accuracy and recall, which is suitable for evaluating the overall performance of classification tasks. The results are shown in **Table 9**.

Table 9. F1-score comparison results (by Category).

Data Category	**ViT (Transformer) (%)**	**ResNet-50 (%)**	**DenseNet-121 (%)**
Normal Bone	94.4	90.6	91.3
Fracture Region	90.1	85.6	86.7
Arthritis Wear	88.6	84.8	85.8
Soft Tissue Lesion	91.7	87.7	88.8
Joint Replacement	88.9	85.7	86.6

ViT's Overall Performance: ViT achieves the highest F1-scores across all categories, particularly excelling in "Normal Bone" (94.4%) and "Soft Tissue Lesion" (91.7%).

CNN Models' Limitations: ResNet-50 and DenseNet-121 show relatively lower F1-scores in complex categories such as "Fracture Region" and "Arthritis Wear", indicating a weaker balance between precision and recall.

ViT's Advantage: ViT's ability to combine high precision and recall through global feature modeling ensures a superior balance, making it especially effective for complex and fine-grained classifications.

Thus, a variation diagram as shown in **Figure 4** can be drawn.

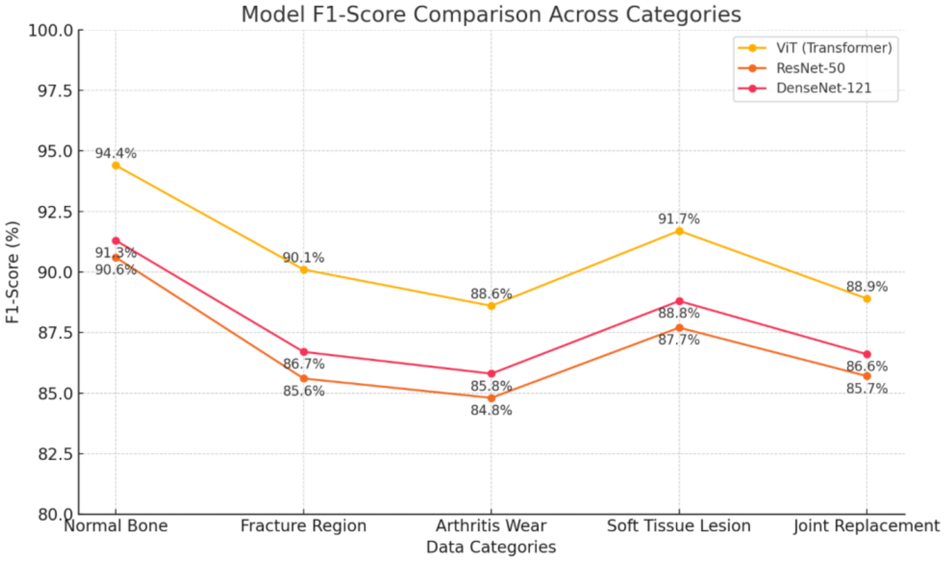

Figure 4. Model F1-score comparison across categories.

In addition to comparing the ViT model with traditional CNNs (such as ResNet-50 and DenseNet-121), we also compared ViT with other state-of-the-art architectures that have shown promise in image recognition tasks, such as EfficientNet, Swin Transformer, and ConvNeXt. The experimental results are shown below (see **Table 10**):

Table 10. F1-score comparison with state-of-the-art techniques.

Model	**F1-Score (%)**
ViT (Transformer)	92.3
EfficientNet	90.4
Swin Transformer	91
ConvNeXt	90.7
ResNet-50	86.9
DenseNet-121	87.7

From **Table 10**, it can be seen that ViT still has a significant advantage in F1-score, particularly in complex structures and multi-class tasks, where ViT shows outstanding performance. Although Swin Transformer (91.0%) and EfficientNet (90.4%) are close to ViT in F1-score, their performance is slightly lower, possibly due to the localized self-attention mechanism of Swin Transformer, which limits global information modeling, and EfficientNet's focus on optimizing computational efficiency, which may sacrifice some classification performance.

ConvNeXt (90.7%) performs well in F1-score but still slightly lags behind ViT. This suggests that ConvNeXt, while optimized based on the convolutional neural network architecture, still has an advantage, but it is not as effective in capturing long-range dependencies and global features compared to the Transformer architecture.

5.5. Confusion matrix analysis

Confusion matrix can visually show the classification results of the model, indicating the correct classification and misclassification of various categories. (see **Table 11**)

Table 11. Confusion matrix of the ViT model.

Predicted/True	**Normal Bone**	**Fracture Region**	**Arthritis Wear**	**Soft Tissue Lesion**	**Joint Replacement**
Normal Bone	282	3	4	2	9
Fracture Region	5	181	7	4	3
Arthritis Wear	6	8	134	5	3
Soft Tissue Lesion	3	4	6	184	3
Joint Replacement	7	5	4	6	128

Normal Bone and Soft Tissue Lesion exhibit the lowest error rates, with most predictions concentrated on their true categories.

Arthritis Wear and Fracture Region show partial misclassification, reflecting the structural similarity in their features, which makes distinguishing between these categories more challenging.

Overall, the ViT model demonstrates high accuracy and robustness across most categories. The relatively low error rates indicate its superior ability to model global features and handle subtle variations.

Thus, a variation diagram as shown in **Figure 5** can be drawn.

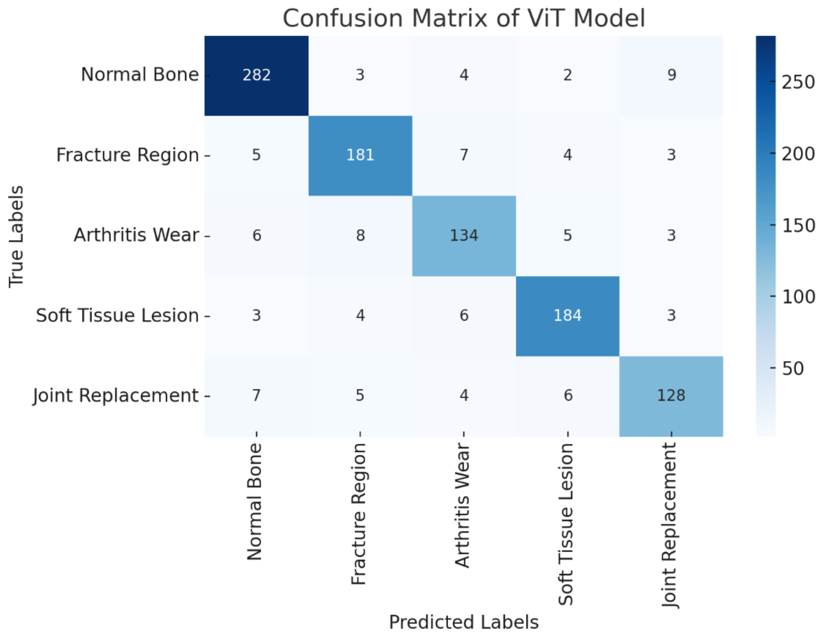

Figure 5. Confusion matrix of ViT model.

6. Discussion

6.1. Research summary

Focusing on "Application of Deep Learning in Biomechanical Image Recognition: Taking Transformer Architecture as the Core", this study proposes a deep learning model based on Vision Transformer (ViT) for the complex structure and recognition challenges of biomechanical images, and makes systematic experimental verification and performance evaluation.

ViT model is proposed and applied to segment the input biomechanical images by Patch and model the self-attention mechanism, which effectively realizes the global feature extraction. Compared with traditional convolutional neural networks (ResNet-50, DenseNet-121), ViT model shows remarkable advantages in identifying complex biomechanical images.

Through the accuracy, precision, recall, F1 score and other indicators, this paper makes a comprehensive comparative analysis between the ViT model and the classic CNN model. The results show that the ViT model is superior to the comparative model in all evaluation indicators. The accuracy rate of ViT model is 92.3%, and the accuracy rate, recall rate and F1 score are balanced and excellent in all categories, especially in the categories of "normal bones" and "soft tissue lesions".

Experiments show that ViT model can effectively capture complex structures and subtle differences in biomechanical images through global feature modeling and multi-head self-attention mechanism, and significantly improve recognition accuracy and robustness. This study provides a new method and technical support for the task of biomechanical image recognition, and an effective tool for clinical diagnosis, disease screening and surgical planning.

6.2. Future work

Although this research has made some progress, there are still some shortcomings and challenges in practical application. Future research will further deepen the following aspects:

Although the Transformer architecture performs well in dealing with global information, its high computational complexity leads to long training time when dealing with large-scale image data, and its application in medical image processing may be limited. The future work will focus on the exploration of lightweight Transformer model. We plan to reduce the parameters and calculation burden of the model by model pruning, knowledge distillation and other technologies, while maintaining its good performance. In addition, considering the complexity of medical images, we will also introduce a hybrid architecture, combining Transformer with Convolutional Neural Network (CNN). CNN can effectively extract local features, while Transformer can carry out global modeling. This combination is expected to improve the reasoning speed and the accuracy of the model.

The data set used in this study is small in scale and covers limited categories, which poses a certain challenge to the generalization ability of the model. Future work will introduce a larger set of biomechanical image data, aiming at covering more disease categories and more complex biomechanical features. This will help to improve the generalization ability of the model and make it adapt to a wider range of clinical application scenarios. We plan to collect data from many hospitals and research institutions to ensure the diversity and representativeness of the data. At the same time, in order to avoid the influence of category imbalance on model training, technologies such as balanced sampling or weighted loss function will be adopted.

With the development of 3D medical imaging technology, future research will explore the application of Transformer in 3D medical images (such as 3D CT and MRI data). We plan to develop a Transformer model for 3D structures, aiming at more accurate 3D anatomical recognition. Specifically, in the future, we will try to fuse the 3D Convolution Network (3D CNN) and Transformer to capture the spatial relationship in the 3D structure. This challenge involves not only the design of model architecture, but also how to process huge 3D image data efficiently.

In the medical field, the interpretability of deep learning model is very important, especially in clinical application, doctors and medical personnel need to understand the decision-making process of the model. In order to improve the interpretability of the model, future research will combine visualization techniques, such as Grad-CAM and Class Activation Mapping (CAM), to analyze the decision-making process of Transformer model in the process of biomechanical image recognition. This not only helps to improve the transparency of the model, but also helps doctors to understand the prediction basis of the model, thus improving the trust and acceptance of users.

Biomechanical images are usually highly complementary to other types of data, and these multimodal data can provide more comprehensive patient information, thus improving the diagnostic accuracy. Future work will explore the application of Transformer in multimodal data fusion, and study how to effectively combine data from different sources (such as image data and clinical text data). We plan to use the multi-modal learning method in deep learning to input different types of data into a unified model framework, and further improve the applicability and diagnostic effect of the model in the actual clinical environment.

Author contributions: Conceptualization, ZY and WF; methodology, ZY; software, ZY; validation, WF; formal analysis, ZY; investigation, WF; resources, WF; data curation, ZY; writing—original draft preparation, ZY; writing—review and editing, WF; visualization, ZY; supervision, ZY. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Yu X, Li S, Zhang Y. Incorporating convolutional and transformer architectures to enhance semantic segmentation of fine-resolution urban images. European Journal of Remote Sensing. 2024; 57(1). doi: 10.1080/22797254.2024.2361768
2. Garibaldi-Márquez F, Martínez-Barba DA, Montañez-Franco LE, et al. Enhancing site-specific weed detection using deep learning transformer architectures. Crop Protection. 2025; 190: 107075. doi: 10.1016/j.cropro.2024.107075
3. Huang C, Wu Z, Xi H, et al. kMaXU: Medical image segmentation U-Net with k-means Mask Transformer and contrastive cluster assignment. Pattern Recognition. 2025; 161: 111274. doi: 10.1016/j.patcog.2024.111274

4. Song Y, Zhou Q. Bi-Modal Bi-Task Emotion Recognition Based on Transformer Architecture. Applied Artificial Intelligence. 2024; 38(1). doi: 10.1080/08839514.2024.2356992
5. Asha CS, Siddiq AB, Akthar R, et al. ODD-Net: a hybrid deep learning architecture for image dehazing. Scientific Reports. 2024; 14(1). doi: 10.1038/s41598-024-82558-6
6. Guerrero-Pantoja D, Pautsch E, Almeida C, et al. Accelerating uncertainty methods for distributed deep learning on novel architectures. The Journal of Supercomputing. 2024; 81(1). doi: 10.1007/s11227-024-06818-y
7. Zhang Z, Song W, Wu Q, et al. A novel local enhanced channel self-attention based on Transformer for industrial remaining useful life prediction. Engineering Applications of Artificial Intelligence. 2025; 141: 109815. doi: 10.1016/j.engappai.2024.109815
8. Paul A, Mallidi S. Enhancing signal-to-noise ratio in real-time LED-based photoacoustic imaging: A comparative study of CNN-based deep learning architectures. Photoacoustics. 2025; 41: 100674. doi: 10.1016/j.pacs.2024.100674
9. Arumai Shiney S, Geetha R. AGBUNet: an enhanced CNN-UNET architecture for the prediction of above ground biomass using deep learning. Neural Computing and Applications; 2024.
10. Omari Alaoui A, Boutahir MK, El Bahi O, et al. Accelerating deep learning model development—towards scalable automated architecture generation for optimal model design. Multimedia Tools and Applications; 2024.
11. Tanneeru VR, Miriyala S, Narukull VR, et al. A deep learning model employing Bi-LSTM architecture to predict Martian ionosphere electron density using data from the Mars Global Surveyor mission. Advances in Space Research. 2024; 74(12): 6343-6355. doi: 10.1016/j.asr.2024.07.051
12. Yang Z, Li G, Xue G, et al. A novel multi-sensor local and global feature fusion architecture based on multi-sensor sparse Transformer for intelligent fault diagnosis. Mechanical Systems and Signal Processing. 2025; 224: 112188. doi: 10.1016/j.ymssp.2024.112188
13. Hu Z, Wang Y, Qi H, et al. Real-time 3D temperature field reconstruction for aluminum alloy forging die using Swin Transformer integrated deep learning framework. Applied Thermal Engineering. 2025; 260: 125033. doi: 10.1016/j.applthermaleng.2024.125033
14. Wang Z, Wang B, Dou H, et al. Windows deep transformer Q-networks: an extended variance reduction architecture for partially observable reinforcement learning. Research square; 2024.
15. Naumova K, Devos A, Karimireddy SP, et al. MyThisYourThat for interpretable identification of systematic bias in federated learning for biomedical images. npj Digital Medicine. 2024; 7(1). doi: 10.1038/s41746-024-01226-1
16. Ilyasova NYu, Demin NS. Systems for Recognition and Intelligent Analysis of Biomedical Images. Pattern Recognition and Image Analysis. 2023; 33(4): 1142-1167. doi: 10.1134/s105466182304020x
17. Qiao M, Dong S. Analysis of Biomechanical Parameters of Martial Arts Routine Athletes’ Jumping Difficulty Based on Image Recognition. Computational intelligence and neuroscience; 2024.