BMC Medical Informatics and Decision Making

RESEARCH Open Access

Feasibility of YOLOX computer model-based assessment of knee function compared with manual assessment for people with severe knee osteoarthritis

Tao Yang[1,2†], Jie Zhao[3†], Ben Wang[2,4], Li Wang[2,5], Hengzhe Bao[2,6], Bing Li[1], Wen Luo[1], Huiwen Zhao[1] and Jun Liu[1*]

Abstract

Objective This study aimed to assess the feasibility of computer model-based evaluation of knee joint functional capacity in comparison with manual assessment.

Methods This study consisted of two phases: (1) developing an automatic knee joint action recognition and classification system on the basis of improved YOLOX and (2) analyzing the feasibility of assessment by the software system and doctors, identifying the knee joint function of patients, and determining the accuracy of the software system. We collected 40–50 samples for use in clinical experiments. The datasets used in this study were collected from patients admitted to the Joint Surgery Center. In this study, the knee joint assessment items included stair climbing, walking on uneven surfaces, and knee joint function. To assess the computer model's automatic evaluation of knee joint function, MedCalc 20 statistical software was used to analyze the consistency of the Lequesne functional index between the computer model's automated determinations and manual independent assessments.

Results The weighted kappa coefficients between the doctors' assessments and the software system's assessments were 0.76 (95% confidence intervals:0.59 ~ 0.92) for climbing up and down stairs, 0.64 (95% confidence intervals:0.45 ~ 0.82) for walking on uneven floors, and 0.68 (95% confidence intervals:0.53 ~ 0.84) for the Lequesne functional index, indicating good consistency between the assessments of the software system and doctors.

Conclusion This paper introduces an automatic knee joint action recognition and classification method based on improved YOLOX. By comparing the results obtained by orthopedic doctors and the software system, the feasibility of this software system was validated in the clinic.

Keywords Deep learning, Knee, Function, Assessment

†Tao Yang and Jie Zhao contributed equally to this work.

*Correspondence:
Jun Liu
drliujun1968@126.com
[1]Joint Surgery Department, Tianjin Hospital, No. 406, Jiefangnan Road, Tianjin 300211, People's Republic of China
[2]College of Orthopedics, Tianjin Medical University, Tianjin, People's Republic of China
[3]Orthopedics Department, Tianjin Hospital, Tianjin, People's Republic of China
[4]Orthopedics Department, Tianjin Occupational Disease Prevention Hospital, Tianjin, People's Republic of China
[5]Orthopedics Department, The Third Hospital of Baotou City, Baotou, People's Republic of China
[6]Orthopedics Department, School of Medicine, Tianjin First Central Hospital, Nankai University, Tianjin, China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Accurate assessment of knee function in patients with knee osteoarthritis is important for guiding treatment decisions and monitoring the natural history of physical dysfunction and osteoarthritis associated with this disease [1–3]. Joint-specific questionnaires such as the American Knee Society Score (AKSS) and the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) have been specifically validated for use in patients with knee osteoarthritis [4]. While these questionnaires help capture symptoms and functional limitations that are important to the patient, they may not provide an accurate representation of underlying joint health or actual functional ability [5]. When questionnaires inquire about pain and difficulty with tasks, there is no standardization of how these tasks are performed. Furthermore, responses to questionnaires can be influenced by other psychosocial factors unrelated to the injured joint, including depression, fatigue, and cognition [6]. Knee function tests in patients with knee osteoarthritis can be feasibly administered preoperatively, which may provide additional longitudinal information about knee function that complements patient questionnaires.

Patient-reported measures of knee function are important for the comprehensive assessment of knee joint disease in both clinical and research contexts [7]. The dimensions that have been deemed important to patients include pain, function, quality of life, and activity level. Artificial intelligence (AI) may afford the opportunity for observations that lead to a new understanding and improved knee function measurement [8]. AI for health care has already provided new perspectives on automated assessments, leading to novel and timely interventions [9]. A common limitation in the development of machine learning (ML) health-behavior models is the time needed to provide real-world context (ground truth) for datasets. However, evidence exists that techniques accounting for small samples and low levels of consistent reporting can produce robust models [10].

Object tracking, which is based on deep learning, offers notable advantages in terms of robustness. However, challenges such as occlusions and poor lighting conditions have spurred the development of innovative solutions. The joint detection and embedding (JDE) model has been introduced to merge reidentification and detection branches [11], thereby enhancing the precision of target detection. The You Only Look Once (YOLO) series, derived from the JDE model, encompasses single-stage object detection algorithms. Spanning from YOLO V1 to YOLO V7 and including various refined versions, this series is characterized by efficiency, flexibility, and superior generalization capabilities. YOLOX, which is a progression in the YOLO algorithmic lineage, has evolved to include anchor-free detectors, offering a blend of rapidity and high accuracy [12].

Conventional knee joint functional assessments typically assume that only medical practitioners measure knee joint function, which requires substantial time and effort. Previous researchers have reported a poor correlation between questionnaire data and functional assessments ($P = 0.08 \sim 0.59$), which suggests that patient perception may be distinct from actual joint function [13]. To increase the precision and efficiency of lower limb functional evaluation in patients with knee osteoarthritis, this study employed deep learning techniques to construct an automated knee joint functional scoring and classification model. By harnessing computer vision, this approach automates patient assessment, significantly enhancing the accuracy and efficiency of scoring while alleviating the workload of medical professionals. Moreover, this method allows for a more nuanced assessment of knee joint functionality, thus providing vital insights for the formulation of tailored treatment plans.

Consequently, the proposed method of automated knee joint functional scoring holds wide-ranging research significance and practical utility. It not only advances the precision of medical evaluations but also has potential applications in the clinical rehabilitation and sports training domains, bolstering training efficacy and safety. Thus, the automated knee joint functional scoring approach delineated in this study has considerable implications for both research and practical implementation. Based on the requirements of clinical diagnosis and patient treatment for knee osteoarthritis, this study preliminary analyzed the feasibility of YOLOX computer model-based knee function assessment and its comparison with manual assessment.

Design

Software development

YOLOX model

The YOLOX algorithm, which was introduced by Megagenius Inc. in 2021 as an enhancement of YOLOv3, presents a distinctive set of attributes [12]:

Decoupled Head: A pivotal advancement lies in the decoupling of prediction branches, engendering a notable acceleration in the convergence rate.

Data Augmentation: Mosaic and mixup techniques are strategically implemented. Intriguingly, data augmentation is gradually phased out in the final 15 epochs to mitigate the risk of excessive augmentation.

Anchor Refinement: YOLOX adopts an anchor-free approach, revolutionizing multipositive and SimOTA methodologies. This dual-pronged refinement not only truncates the training time but also markedly increases the predictive accuracy (Fig. 1).

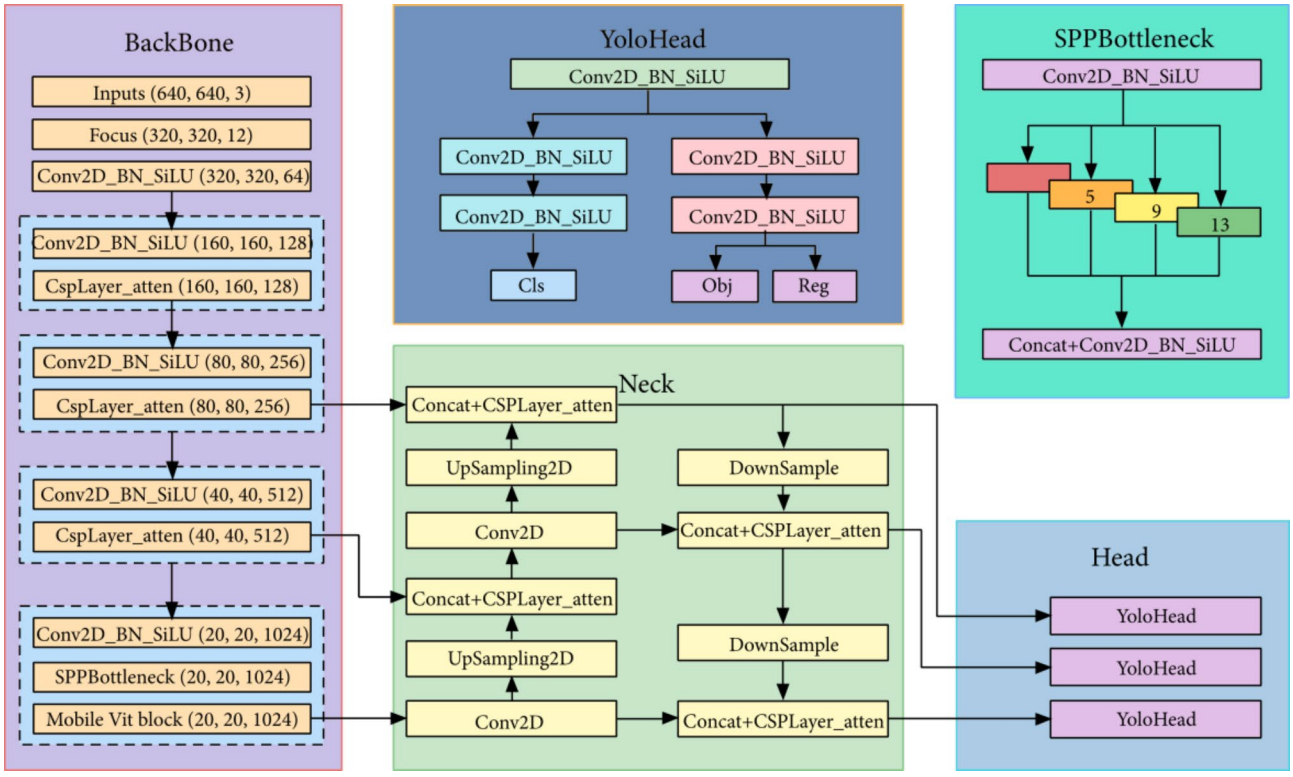

Fig. 1 YOLOX Model Framework

Table 1 Differences in performance parameters of YOLOX-l compared with other models

Model	Input Resolution	Mean average precision (mAP) (%)	Size (MB)	Inference Speed (FPS)	Training Time (hours)
YOLOX-s	640×640	76.39	32.0	52	4
YOLOX-m	640×640	83.69	90.2	43	12
YOLOX-l	640×640	91.73	192.9	28	20

In the standard model of the YOLOX framework, the mean average precision (mAP) of YOLOX-l was 8.8% greater than that of YOLOX-s, as demonstrated in Table 1. The training dataset that was used in this study encompasses data capturing a spectrum of human movements, including climbing up and down stairs, squatting and walking on uneven floors. The strategic inclusion of such diverse actions enables the model to learn from a range of human poses, enhancing the model's robustness and generalization capabilities. Considering the task scenario and requirements combined with the visual view, YOLOX-l, which has high prediction accuracy, was selected.

Bayesian optimization

Bayesian optimization is an optimization approach based on Bayes' theorem, which, within a limited number of iterations, progressively evaluates the objective function to discover the optimal combination of hyperparameters [14]. Compared with traditional methods such as random search and grid search, Bayesian optimization achieves faster convergence to the optimal solution by leveraging prior knowledge and confidence intervals, resulting in higher search efficiency and accuracy. In this study, Bayesian optimization was applied to optimize the following hyperparameters:

Learning Rate: The learning rate plays a vital role in controlling the magnitude of model weight updates in deep learning. It is typically represented as η or α. The magnitude of the learning rate directly influences the model's training effectiveness.

Batch Size: The batch size refers to the number of samples used in each iteration when training a neural network.

Exponential Moving Average Decay: Exponential moving average decay is a commonly used optimization technique in deep learning. It is employed primarily to smooth the variations in weights within a model and reduce the volatility of weight updates, enhancing the model's stability. The formula for exponential moving average decay is as follows:

$$\theta_{t+1} = \beta\theta_t + (1-\beta)\theta_{t+1} \tag{1–1}$$

Table 2 The setting intervals and results of bayesian optimization hyperparameters

Hyperparameter	Optimization interval	Optimal Results
Learning rate	[0.001, 0.1]	0.03
Batch size	[16, 64]	32
Sliding average attenuation	[0.9, 0.99]	0.95

where θ_t represents the parameter values before iteration t, θ_{t+1} represents the parameter values after iteration t, and β represents the attenuation factor, whose value typically ranges between 0.9 and 0.999. Moving average attenuation can improve the generalization ability of the model and reduce the risk of overfitting (Table 2).

Improved SCP module

The CSP architecture derives inspiration from the network structure of SCPNet, ingeniously amalgamating convolutional layers and X residual components to augment information propagation and improve feature representation.

The learning potential of convolutional neural networks (CNNs) is attributed to the SCP module, which remains constrained by the dimensions of the convolutional kernel, precluding the holistic integration of pixel-level information beyond the kernel bounds.

In object recognition, the proposed model has better precision than the traditional model does in terms of accuracy and speed. The study takes a holistic perspective, emphasizing the model's efficacy. As a result, this study introduces crucial refinements into the original SCP module. Through a reduction in the core backbone's parameters and the elimination of redundant gradient information during inference, these modifications improve the model's capacity for learning (Fig. 2).

Each group of 3×3 convolutions in the output channels of the original SCP module is preceded by a set of 1×1 convolutions, effectively halving the output channel

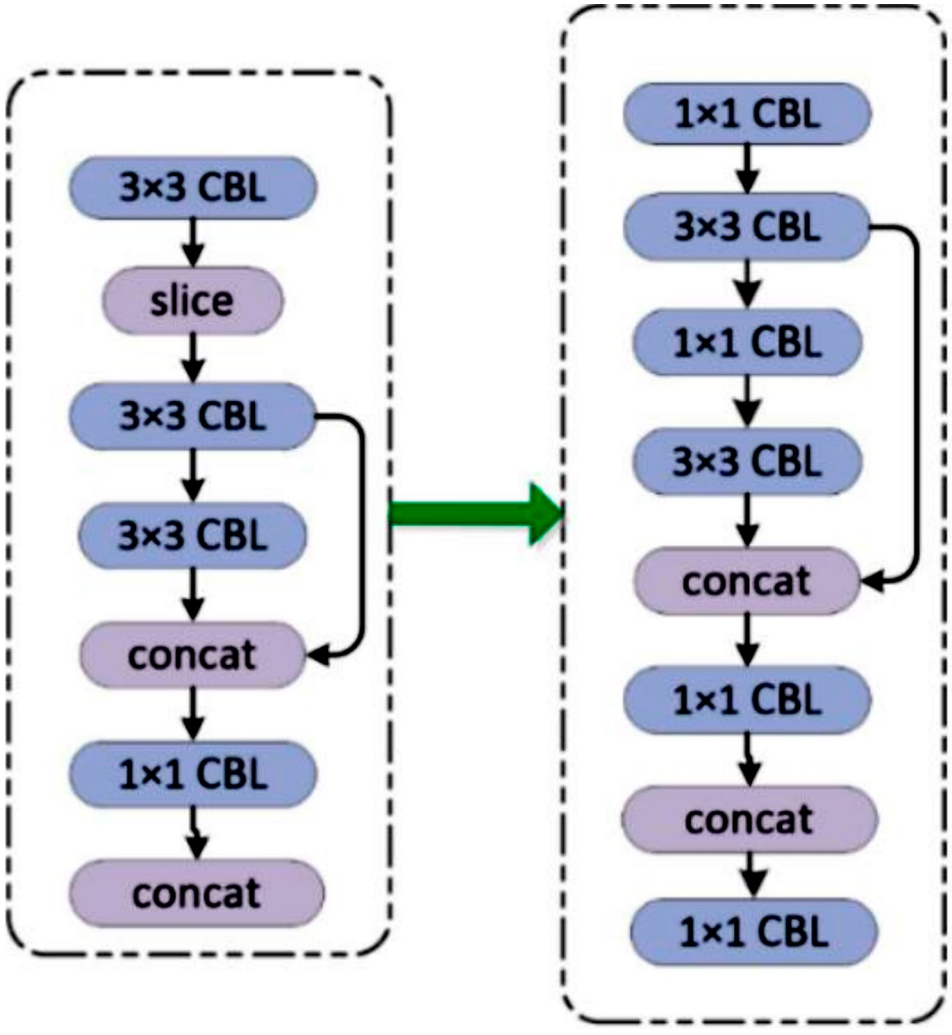

Fig. 2 Improve SCP module

count while maintaining the input channel dimensions. The input feature maps are handled by the 3 × 3 convolutional block residual structure, whereas the output feature maps are processed through the 1 × 1 convolutional block, ensuring dimensional congruence between the input and output feature maps. Ultimately, this innovation enables the incorporation of three residual blocks on the right side of the new module, effectively achieving the same receptive field as the SCP module.

In YOLOX, convolution operations are performed on the parameters of the original SCP module, and the expression is as follows:

$$\begin{aligned} P &= \sum_{I=1}^{N} (C_I\_in * C_I\_out * K_I^2 \\ &= 3*3*C^2 + 3*3*\frac{C}{2}*\frac{C}{2} + 3*3*\frac{C}{2}*\frac{C}{2} + 1*1*C^2 \\ &= 14.5C^2 \end{aligned} \quad (1\text{–}2)$$

The improved expression for parameter calculation in this study is as follows:

$$\begin{aligned} P &= \sum_{I=1}^{N} (C_I\_in * C_I\_out * K_I^2 \\ &= 1^2 * C * \frac{C}{2} + \left((3^2+1^2) * \frac{C}{2} * \frac{C}{2}\right) * 3n + 1^2 \\ &* \frac{C}{2} * \frac{C}{2} + 1^2 * C^2 = 9.25C^2 \end{aligned} \quad (1\text{–}3)$$

where I is the number of layers in the network, C_{l_in} is the number of input channels in layer I, C_{I_out} is the number of output channels in layer I, and K is the size of the convolutional kernels.

In formulas (1–2) and (1–3), when the receptive fields are the same, the computational complexity of the improved structure is reduced by 36.2% compared with that of the original CSP module.

Experimental Setup and Environment: The experimental investigations detailed in this study were conducted within a controlled computational environment. The hardware configuration utilized for experimentation comprised an Intel(R) Core(TM) i9-10900 K CPU, 128 GB of RAM, and an NVIDIA RTX 3090 graphics card with 24 GB of memory. The software infrastructure employed was anchored by the Ubuntu 20.04 LTS 64-bit operating system.

Dataset construction

The dataset utilized in this study was sourced from inpatients at the Joint Surgery Center who were prepared for total knee arthroplasty. The inclusion criteria for patients were as follows: ① undergoing preparation for total knee arthroplasty and ② diagnosed with severe osteoarthritis of the knee, classified as KL grade III or higher through radiological imaging. The exclusion criteria for patients were as follows: ① afflicted by systemic immune disorders, leading to a reduced quality of daily life, such as rheumatoid arthritis, and ② having severe preoperative knee deformities impairing independent mobility.

① Data Augmentation:

Data augmentation is a technique involving the application of transformations and expansions to original data to increase sample diversity, thereby enhancing model generalization performance and mitigating overfitting. With respect to the input requirements of the improved YOLOX, the following data augmentation operations were executed:

a. Random Cropping: Extracting a random portion of an image for training aids the model in learning object features at various positions.
b. Random Flipping: Horizontally or vertically flipping images at random adds diversity to the dataset, enabling the model to adapt to different object orientations.
c. Random Rotation: Introducing random rotations simulates the appearance of objects at varying angles.
d. Random Scaling: Randomly scaling images ensures that objects can be accurately detected at different sizes.
e. Random Brightness and Contrast Adjustment: Randomly adjusting image brightness and contrast increases image variability.

Data augmentation fosters diverse sample representations, enhances model robustness and leads to improved detection performance during training.

② Model Training:

During model training, the annotated dataset was partitioned into training and validation sets. 70% of the data were allocated for training, and the remaining 30% were allocated for validation. This division ensured ample samples for model learning during training and facilitated the assessment of model generalization on the validation set.

The primary challenge of the traditional YOLO loss function lies in unstable IoU calculations when objects exhibit substantial overlap. The CIoU loss function addresses this by considering the comprehensive intersection over union between objects and incorporating distance metrics.

The formula for the CIoU loss function is as follows:

$$\mathrm{CIoU\ Loss} = -\mathrm{IoU} + \frac{d^2}{c^2} \quad (1\text{–}4)$$

where IoU (intersection over union) represents the intersection area of the predicted box and the real box divided by their union area; d^2 is the square of the Euclidean distance between the center points of the prediction box and the real box; and c^2 is a parameter used to punish size differences between the boxes.

Stochastic gradient descent (SGD) is a widely employed optimization algorithm for training neural networks and diverse machine learning models. A first-order optimization algorithm updates model parameters utilizing the gradient of a single sample, eschewing the computation of the average gradient across all samples. This feature not only renders SGD computationally efficient but also makes it suitable for optimizing large-scale datasets.

The formula for SGD is as follows:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla J(\theta_t, x_i, y_i) \tag{1–5}$$

where θ_t is the parameter of the model at the tth iteration and η is the learning rate, which is used to control the step size of each parameter update. It is a hyperparameter and needs to be set beforehand. Too large a learning rate may result in too drastic a parameter update, whereas too small a learning rate may result in too slow convergence. $J(\theta_t, x_i, y_i)$ is the loss function, which represents the prediction error of the model for the training sample (x_i, y_i) under parameter θ_t. By calculating the gradient $\nabla J(\theta_t, x_i, y_i)$ of the loss function with respect to the parameter, SGD aims to minimize the loss function and thus optimize the parameters of the model.

In SGD, each iteration involves the selection of a random training sample (x_i, y_i) to compute the loss function and gradient. Leveraging the gradient's direction and the learning rate, the model parameters are then updated. This process iterates over the entire training dataset multiple times, culminating when the predetermined number of training epochs is achieved or when convergence conditions are met.

In summary, the settings used for model training in this study were as follows:

- Training Set: 70% of the annotated dataset;
- Validation Set: 30% of the annotated dataset;
- Loss Function: the YOLOX object detection loss function (CIoU loss function);
- Optimization Algorithm: stochastic gradient descent (SGD).

③ Evaluating the detection model performance

When assessing the efficacy of an object detection model, the metric of choice is the mean average precision. mAP is derived by calculating precision values across varying confidence thresholds and subsequently computing their average. Several relevant formulas encompass the evaluation process:

Precision signifies the proportion of samples correctly classifies as positive by the model among all samples classified as positive.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1–6}$$

The recall rate represents the proportion of samples that the model successfully predicts to be positive among all true positive samples.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1–7}$$

The F1 score represents the harmonic average of accuracy and recall, which is used to comprehensively consider the accuracy and comprehensiveness of the model.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1–8}$$

mAP represents the average precision value calculated at different confidence thresholds.

The prediction outcomes of the model are meticulously ordered on the basis of confidence scores. By incrementally adjusting the confidence threshold, precision values are subsequently calculated at each threshold, which are then graphed into a precision–recall curve.

④ Model recognition and classification process

a. Region Calibration: Pertinent regions within the test area undergo precise calibration, effectively delimiting the testing scope. This calibration strategy ensures testing uniformity and reproducibility, as depicted in Fig. 3.

In the pursuit of refining object detection evaluation and improving the model recognition process, the assessment methodology hinges upon precise precision–recall calculations. This approach not only offers insights into model performance but also facilitates robust classification in diverse scenarios.

b. System Testing: Upon entering the test area, patients' movements are timed. The walking speed on uneven terrain is calculated using the video frame rate and the number of frames in which individuals are detected. Additionally, the squatting and stair-climbing times are recorded.

Walking Speed Calculation Formula:

Walking Speed = Time Taken / Distance Walked.

Distance Walked: Total distance covered by the patient within the test area, representing the actual range of the designated zone.

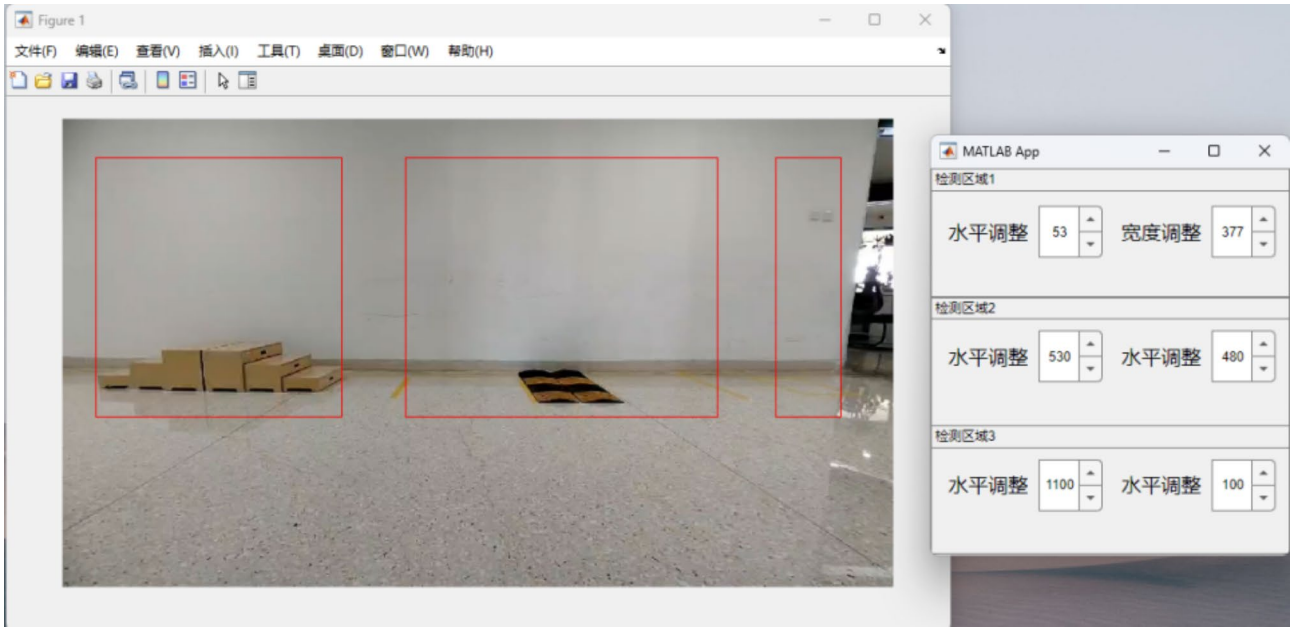

Fig. 3 Model Calibration Area

Time Taken: Obtained by dividing the number of frames capturing the patient's movement by the video frame rate.

Squatting and Stair-Climbing Time Recording:

Squatting Time Calculation: The time taken for a single squatting action by the patient, directly recorded during the test.

Stair-Climbing Time Calculation: The time required by the patient to complete the stair-climbing action, recorded during the test.

Importantly, walking speed is typically measured in distance per unit time (e.g., meters per second), whereas squatting and stair-climbing times are measured in seconds. By calculating patients' walking speed, squatting time, and stair-climbing time, their mobility and performance can be effectively assessed.

c. Calculating Walking Speed: Using the video frame rate and the frames used to detect individuals, the walking speed of each patient on uneven terrain can be calculated. The speed is determined by dividing the total distance walked by the patient within the area by the total time.
d. Calculating Squatting Time: For squatting actions, the time taken for a single squatting action by a patient can be calculated. The squatting action can be defined on the basis of a predetermined threshold for the patient's squatting height. The time when the patient's squat height meets the set threshold is recorded.
e. Calculating Stair-Climbing Time: For stair-climbing, the total time taken for the entire action is recorded, spanning from stepping onto the stairs to completing the descent.

Through these steps, patients' abilities in walking on uneven terrain, squatting, and stair climbing are tested, with the corresponding speed and completion times calculated to evaluate and analyze their mobility. The evaluation system is based on time as the evaluation standard, and the completion time of actions is in accordance with the local standard < Specification for evaluation of criterion-referenced senior functional fitness standards > in Tianjin. A completion time percentile < P25 indicates easy performance of the action, a completion time in the P25 ~ P50 percentile range indicates mild difficulty, a completion time in the P50 ~ P75 percentile range indicates moderate difficulty, and a completion time in the P75 ~ P100 percentile range indicates severe difficulty in completing or inability to complete the required action, resulting in movement interruption. This standard is applicable to the functional evaluation of elderly people in Tianjin [15].

Clinical experiments

Study subjects

For the confirmatory analysis, the ratio between the sample size and the number of items had better to be above 5:1 [16]. The ideal sample size is 10 ~ 25 times the number of items. The Lequesne functional index only has 4 items (climbing upstairs, climbing down stairs, squatting down, and walking on uneven floors). Therefore, 40 ~ 50 samples are collected for use in clinical experiments. The datasets used in this study were collected from

patients admitted to the Joint Surgery Center in Tianjin. This center is one of the largest knee joint disease diagnosis and treatment centers in the local area, with more than 70 fixed beds and treating over 1500 patients with knee osteoarthritis every year. The inclusion criteria for patients were as follows: (1) knee osteoarthritis, graded KLIII or above on the basis of radiographic imaging [17], and (2) the capacity to provide informed consent. The exclusion criteria for patients were as follows: (1) severe deformities affecting knee joint mobility, namely, varus knee deformity > 20°, valgus knee deformity > 20°, or degree of knee flexion contracture > 20° [18]; (2) a history of knee joint trauma within the past 3 months; and (3) neurological disorders impacting movement, such as Alzheimer's disease. This clinical study received ethical approval from the institutional medical ethics committee (2023-YLS-078).

Methodology

Clinical doctors (four in total, each with over five years of clinical experience in joint surgery) underwent training in CKFA usage. Researchers provided a detailed explanation of the computer model report structure to the trainees, along with two demonstration videos showcasing the computer model's functionality. This approach aimed to acquaint the assessors with the computer model's output process without providing any instructions on the video actions throughout the training.

Evaluation Tool: The Lequesne functional index, initially proposed in 1987, was adopted for both manual and computer-based knee joint functional evaluations [19]. This index evaluates disease conditions and joint functions in patients with knee joint osteoarthritis and consists of two parts: osteoarthritis symptoms and daily life functional disabilities. The measurement takes approximately 3–5 min. Li et al. validated the Chinese version of the Lequesne Index, which yielded an interrater reliability coefficient (ICC) of 0.94 [20]. Presently, this index is widely utilized to assess patient disease conditions and perform follow-up evaluations in patients with knee joint osteoarthritis.

The clinical trial was divided into two stages: (1) Proficient orthopedic doctors (four in total, each with over 5 years of clinical experience in joint surgery) evaluated knee joint function through clinical physical examinations, and the Lequesne Functional Index was used to assign scores to patients. (2) Three days after undergoing manual knee joint functional evaluation, the subjects executed designated movements, such as climbing up stairs, squatting, and walking on uneven floors. The trained doctors then utilized the computer software model to automatically generate patient knee joint functional reports.

Statistical analysis

In assessing the computer model's automatic evaluation of knee joint function, MedCalc 20 statistical software was employed to analyze the consistency between the computer model's automated determinations and manual independent assessments via the weighted kappa coefficient, which was used to evaluate the reliability of the computer model software in determining knee joint function [21]. The MedCalc version 20.0 (MedCalc Software, Ostend, Belgium) statistical software package was used for statistical analysis of all the variables. The agreement between the clinical doctors' assessments and model classification results was evaluated via Bland–Altman analysis [22]. It was validated through 95% confidence intervals for both the kappa and Bland–Altman assessments.

Results

Patients

A total of 42 patients were included in this study, with 20 males and 22 females. The mean age was 73.6 ± 8.21 years. All the subjects successfully completed two knee joint functional evaluations, and no accidental incidents, such as falls, occurred.

Climbing up and down stairs

Both the clinical doctors and the computer model included in this study independently measured the ability to climb up and down stairs. A consistency evaluation via MedCalc 20 yielded a weighted kappa coefficient of 0.76 and a 95% confidence interval of 0.59–0.92, indicating good agreement between the two methods. In this study, we conducted a Bland–Altman analysis, which revealed that the limits of agreement between manual assessment and model evaluation were notably tight, ranging from − 1.1 to 1.9 (Table 3; Fig. 4).

Walking on uneven floors

In addition to stair-climbing ability, the ability to walk on uneven floors was assessed by both doctors and the computer model in this study. The assessment results were subjected to consistency analysis via MedCalc 20, which yielded a weighted kappa coefficient of 0.64 and a 95% confidence interval of 0.45–0.82 between the two assessment methods. The results revealed that the limits of agreement between manual assessment and model evaluation were notably confined, ranging from − 2.5 to 2.2 (as shown in Table 4; Fig. 5).

Knee joint function

Finally, the knee joint function of the participants was assessed by doctors and the computer model. A consistency analysis was performed via MedCalc 20, resulting in a weighted kappa coefficient of 0.68 and 95%

Table 3 Consistency between manual evaluation and model evaluation of up and down stairs(Kappa)

model evaluation	manual evaluation					
	0	**1**	**2**	**3**	**4**	
0	26	1	1	1	0	29 (69.0%)
1	0	3	0	3	0	6 (14.3%)
2	0	0	1	2	0	3 (7.1%)
3	0	0	0	1	3	4 (9.5%)
4	0	0	0	0	0	0 (0.0%)
	26 (61.9%)	4 (9.5%)	2 (4.8%)	7 (16.7%)	3 (7.1%)	42
Weighted Kappa [a]	0.76					
Standard error	0.08					
95% CI	0.59–0.92					

[a] Quadratic weighting

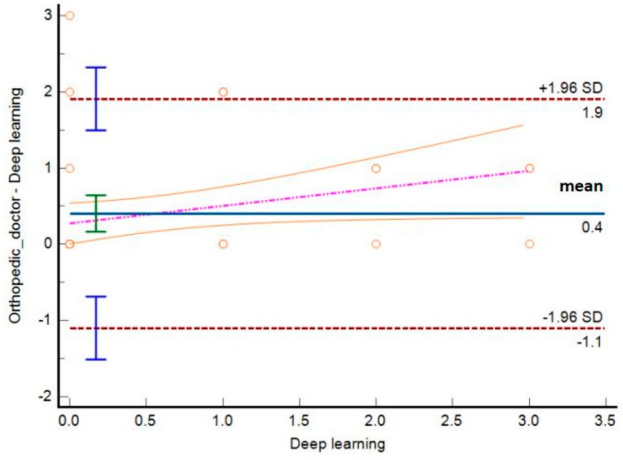

Fig. 4 Individual agreement of up and down stairs between manual evaluation and model evaluation

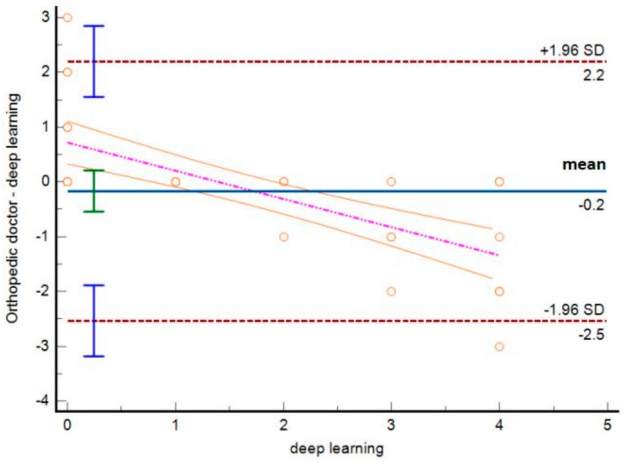

Fig. 5 Individual agreement of walking on uneven floor between manual evaluation and model evaluation

confidence interval of 0.53–0.84 between the two assessment methods. This indicates an acceptable level of agreement (as presented in Table 5). Moreover, through our investigation, we conducted a Bland–Altman analysis, reaffirming that the agreement between manual assessment and model evaluation was notably bounded, ranging from – 2.9 to 3.5 (as depicted in Fig. 6).

Discussion

This study presents the first application of the YOLOX network model to knee joint motion assessment, establishing an automatic classification system for knee joint actions on the basis of the YOLOX network. This system was employed to evaluate patients' daily knee joint functional capacity. Through clinical experiments, we observed that the knee joint action classification system

Table 4 Consistency between manual evaluation and model evaluation of walking on uneven floor(Kappa)

model evaluation	manual evaluation					
	0	**1**	**2**	**3**	**4**	
0	8	3	3	1	0	15 (35.7%)
1	0	8	0	0	0	8 (19.0%)
2	0	1	3	0	0	4 (9.5%)
3	0	1	2	1	0	4 (9.5%)
4	0	1	5	1	4	11 (26.2%)
	8(19.0%)	14(33.3%)	13(31.0%)	3(7.1%)	4(9.5%)	42
Weighted Kappa [a]	0.64					
Standard error	0.09					
95% CI	0.45–0.82					

[a] Quadratic weighting

Table 5 Consistency between manual evaluation and model evaluation of knee joint function(Kappa)

model evaluation	manual evaluation									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
0	2	0	3	1	0	0	0	0	0	6 (14.3%)
1	0	2	0	1	0	0	0	0	0	3 (7.1%)
2	0	0	4	1	1	0	0	0	0	6 (14.3%)
3	0	0	0	3	3	1	1	0	0	8 (19.0%)
4	0	0	3	1	2	0	2	0	0	8 (19.0%)
5	0	0	0	0	0	2	1	1	0	4 (9.5%)
6	0	0	0	0	1	0	1	0	1	3 (7.1%)
7	0	0	0	0	0	0	0	2	0	2 (4.8%)
8	0	0	0	1	0	1	0	0	0	2 (4.8%)
	2(4.8%)	2(4.8%)	10(23.8%)	8(19.0%)	7(16.7%)	4(9.5%)	5(11.9%)	3(7.1%)	1(2.4%)	42
Weighted Kappa [a]	0.68									
Standard error	0.07									
95% CI	0.53–0.84									

[a] Quadratic weighting

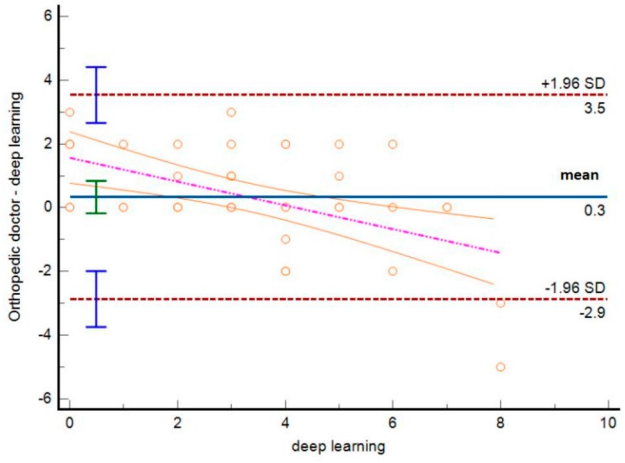

Fig. 6 Individual agreement of knee joint function between manual evaluation and model evaluation

built upon the YOLOX network exhibited certain similarities in assessment outcomes compared with experienced medical practitioners, indicating its potential for use in clinical research.

By assessing and classifying various knee joint actions, we found that the computer model's evaluation results were highly consistent with those of clinical orthopedic physicians in the assessment of stair climbing ability. Action recognition and motion analysis in the medical field have relied mainly on expensive equipment, such as high-speed cameras, Kinect cameras, and optical motion capture devices, and have been limited by other factors, such as scene characteristics and costs [23–25]. We performed action recognition analysis via a computer model. Most research has focused on rehabilitation scenarios [26–29]. Recently, researchers developed a mobile app based on the AlphaPose and VideoPose algorithms for evaluating patients' knee joint function and stiffness by recording five sit-to-stand test videos [30], whose technology uses self-assessment by patients and simple application scenarios. Compared with several studies [31–34], our research revealed that computer models often demonstrate greater accuracy in assessing actions with larger ranges of motion. This phenomenon may be attributed to several factors: (1) Actions with larger knee joint ranges of motion (such as stair climbing and squatting) tend to have distinctive features, enabling the computer model to rapidly identify and classify them. (2) Many patients with severe knee joint osteoarthritis exhibit limited joint mobility and require maximum effort to complete actions with larger ranges of motion, resulting in slower movements that are more easily evaluated by clinical physicians. (3) Relative to stair climbing, the assessment of walking on uneven surfaces is simpler for patients with severe knee joint osteoarthritis, requiring lower joint mobility. Consequently, patients complete the action in a shorter time during evaluation, making discrepancies more likely to arise between manual and computer model assessments. (4) During manual knee joint functional assessments, clinical physicians often consider factors such as overall physical function (e.g., cardiorespiratory fitness) and mental state (e.g., fatigue) to derive final assessment outcomes. In contrast, computer models evaluate knee joint function solely on the basis of the completion of actions without incorporating other factors, such as psychological state and environmental factors.

In the past, many studies achieved action recognition and behavior assessment through OpenPose systems [35]. In contrast to other automated classification assessment systems for rehabilitation and sports [36], our study first developed a knee joint action automatic classification evaluation system based on the enhanced YOLOX model. Compared with common human pose recognition algorithms such as OpenPose, the YOLOX

algorithm has the advantages of low computational complexity, strong generalization ability, good scene adaptability and robustness [37]. Moreover, research has confirmed that establishing a computer vision system with algorithmic services is crucial for action assessment and behavior prediction [38]. This system employs Bayesian hyperparameter optimization for loss function improvement, expediting model convergence and reducing the computational burden on servers and GPUs, thus increasing computational efficiency. By incorporating the Lequesne knee joint functional index and addressing the specific requirements and application scenarios of clinical physicians and rehabilitation trainers, our research modularized the design and business logic, including action completion time and assessment score reports. This was followed by algorithm deployment into the knee joint motion functional classification system. Sequential orchestration of the YOLOX model's action recognition and serialization services provided real-time action scoring for participants.

While this study successfully applied the YOLOX model to knee joint action recognition and conducted preliminary clinical tests that demonstrated certain similarities with orthopedic physicians' manual assessments, several limitations remain. These include the following: (1) The clinical trial's sample size was limited due to factors such as patient demographics, hospital environmental factors, and research project duration, reducing the persuasiveness of the study results. Future work should involve large-scale, multicenter clinical trials to evaluate the effectiveness of the knee joint action recognition and classification system developed in this study for clinical practice. (2) The system requires a high level of computer hardware configuration, with a CPU of i9-10 or above and an independent graphics card, to ensure detection and transmission speed. Otherwise, problems such as delayed operation and low efficiency of the automatic system may occur. (3) The automatic evaluation system developed in this study needs to capture the video information of the patient completing specific movements and evaluate knee function on the basis of an analysis of the video information. The system cannot conduct continuous automatic evaluation. (4) Current research has focused mainly on the initial development and validation stage of the system. This automated scoring system has not been used to conduct a comprehensive evaluation of long-term performance and sustainability in clinical practice. We will carry out long-term observations in real clinical environments to validate the durability and sustained effectiveness of the system in future research. (5) Environmental factors such as lighting conditions and background clutter may indeed affect the model's ability to detect and classify knee joint movements accurately in actual environments. To improve accuracy, we recommend video capture under good lighting conditions and against a simple background. (6) With ongoing technological advancements, computer classification model algorithms continue to evolve. Thus, the project team should further enhance the model algorithm to improve work efficiency and meet clinical requirements. (7) The later-stage scoring module of the project incorporates only a few actions, limiting the generalizability of the model's assessment results.

Conclusions

This paper introduces an automatic knee joint action recognition and classification method based on improved YOLOX, aimed at addressing the challenges of low computational efficiency and limited robustness in computer vision action recognition. A comparison of the measurement results with those of experienced medical practitioners preliminarily verified the potential application of this technology in knee joint function assessment scenarios.

Abbreviations

AKSS	the American Knee Society Score
WOMAC	the Western Ontario and McMaster Universities Osteoarthritis Index
AI	artificial intelligence
ML	machine learning
JDE	Jointly learns the detector and embedding model
YOLO	You Only Look Once
mAP	mean average precision
CNNs	Convolutional Neural Networks
SGD	Stochastic Gradient Descent
ICC	inter-rater reliability coefficient

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02877-0.

Supplementary Material 1

Acknowledgements

The authors are grateful to professor Jin Shoufeng, Xi'an Polytechnic University for guiding technological work.

Author contributions

Jun Liu, Bing Li and Tao Yang wrote the main manuscript text. Jie Zhao, Ben Wang and Li Wang prepared figures. Wen Luo and Huiwen Zhao prepared tables. All authors reviewed the manuscript.

Funding

This study was supported by many institutions, specifically the following: (1) the Tianjin Science and Technology Commission (TSTC) Diversified Fund (No: 21JCZDJC01000; PI: Pro. Jun Liu); (2) the National Natural Science Foundation of China (No: 82102639; PI: Dr. Jie Zhao); and (3) the Tianjin Natural Science Foundation (No. 22JCQNJC00850; PI: Dr. Jie Zhao).
The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

The data are not publicly available owing to restrictions, their containing information could compromise the privacy of research participants.

Declarations

Ethics approval and consent to participate

The current study was approved by the Institutional Ethical Review Board of Tianjin Hospital (No. TJYY-2023-YLS-078) and was conducted in accordance with the Ethical Guidelines for Epidemiological Research by the Chinese Government and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All study participants provided written informed consent by the completion and submission of the survey.

Consent to publish

Before the test video was recorded, the project team obtained patient permission to publish knee joint function test images publicly.

Declaration of the use of generative AI and AI-assisted technologies in the writing process

We did not use the generative AI and AI-assisted technologies to write the manuscript. Generative AI and AI-assisted technologies were only used to check grammar and spelling errors after writing the manuscript.After using this service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Competing interests

The authors declare no competing interests.

Received: 2 November 2023 / Accepted: 17 January 2025
Published online: 24 January 2025

References

1. Becker R, Berth A, Nehring M, et al. Neuromuscular quadriceps dysfunction prior to osteoarthritis of the knee. J Orthop Res. 2004;22:768–73.
2. Bremander AB, Dahl LL, Roos EM. Validity and reliability of functional performance tests in meniscectomized patients with or without knee osteoarthritis. Scand J Med Sci Sports. 2007;17:120–7.
3. Ericsson YB, Roos EM, Dahlberg L. Muscle strength, functional performance, and self-reported outcomes four years after arthroscopic partial meniscectomy in middle-aged patients. Arthritis Rheum. 2006;55:946–52.
4. Irrgang JJ, Anderson AF, Boland AL, et al. Development and validation of the international knee documentation committee subjective knee form. Am J Sports Med. 2001;29:600–13.
5. Thorlund JB, Aagaard P, Roos EM. Thigh muscle strength, functional capacity, and self-reported function in patients at high risk of knee osteoarthritis compared with controls. Arthritis Care Res (Hoboken). 2010;62:1244–51.
6. Roos EM, Bremander AB, Englund M, et al. Change in self-reported outcomes and objective physical function over 7 years in middle-aged subjects with or at high risk of knee osteoarthritis. Ann Rheum Dis. 2008;67:505–10.
7. Padua L, Evoli A, Aprile I, et al. Myasthenia gravis outcome measure: development and validation of a disease-specifific self-administered questionnaire. Neurol Sci. 2002;23:59–68.
8. Tiulpin A, Thevenot J, Rahtu E, et al. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep. 2018;8(1):1727.
9. He K, Zhang X, Ren S et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 770–77.
10. Esteva A, et al. Dermatologist-level classifcation of skin cancer with deep neural networks. Nature. 2017;542:115–8.
11. Zhang YF, Wang CY, Wang XG, et al. Fairmot: on the fairness of detection and re-identification in multiple object tracking. Int J Comput Vision. 2021;129(11):3069–87.
12. Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. Computer Vision and Pattern Recognition. (2018-04-08) [2021-12-15]. https://pjreddie.com/media/files/papers/YOLOv3.pdf
13. Naimark MB, Kegel G, O'Donnell T, Lavigne S, Heveran C, Crawford DC. Knee function Assessment in patients with Meniscus Injury: a preliminary study of reproducibility, response to treatment, and correlation with patient-reported questionnaire outcomes. Orthop J Sports Med. 2014;2(9):2325967114550987.
14. Zhu W, Qu JY, Wu RB. Straight convolutional neural networks Algorithm based on batch normalization for image classification. J Computer-Aided Design&Computer Graphics. 2017;29(9):1650–7.
15. Huang LP, Cao LJ, Liu YG et al. Specification for evaluation of criterion-referenced senior functional fitness standards. DB12/T 1279–2023. https://std.samr.gov.cn/db/search/stdDBDetailed?id=0D14A28AB103BB28E06397BE0A0A4CE3
16. McDonald RP, Comments on DJ, Bartholomew. Foundations of factor analysis: some practical implications. Brit J Math Stat Psychol. 1985;38:134–7.
17. Kellgren JH, Lawrence JS. Radiological assessment of osteoarthrosis. Ann Rheum Dis. 1957;16(4):494–502.
18. Martimbianco AL, Calabrese FR, Iha LA, Petrilli M, Lira Neto O, Carneiro Filho M. Reliability of the American Knee Society Score (AKSS). Acta Ortop Bras. 2012;20(1):34–8.
19. Dawson J, Linsell L, Doll H, et al. Assessment of the Lequesne index of severity for osteoarthritis of the hip in an elderly population. Osteoarthritis Cartilage. 2005;13(10):854–60.
20. Li CHG, YMi,Chen W, et al. Assessment of Interrater reliability of Lequesne Index (Chinese Version) in knee osteoarthritis. Chin J Rehabil Theory Pract. 2010;16(6):554–5.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
22. Bunce C. Correlation,agreement and Bland-Altman analysis: statistical analysis of method comparison studies. Am J Ophthalmol. 2009;148:4–6.
23. Brambilla C, Marani R, Romeo L, Lavit Nicora M, Storm FA, Reni G, Malosio M, D'Orazio T, Scano A. Azure Kinect performance evaluation for human motion and upper limb biomechanical analysis. Heliyon. 2023;9(11):e21606.
24. Milosevic B, Leardini A, Farella E. Kinect and wearable inertial sensors for motor rehabilitation programs at home: state of the art and an experimental comparison. Biomed Eng Online. 2020;19(1):25.
25. Sun J, Liu Y, Yan S, et al. Clinical gait evaluation of patients with knee osteoarthritis. Gait & Posture. 2017;58:319–24. https://doi.org/10.1016/j.gaitpost.2017.08.009.
26. Sardari S, Sharifzadeh S, Daneshkhah A, Nakisa B, Loke SW, Palade V, Duncan MJ. Artificial Intelligence for skeleton-based physical rehabilitation action evaluation: a systematic review. Comput Biol Med. 2023;158:106835.
27. Sardari S, Sharifzadeh S, Daneshkhah A, Loke SW, Palade V, Duncan MJ, Nakisa B. LightPRA: a lightweight temporal Convolutional Network for Automatic Physical Rehabilitation Exercise Assessment. Comput Biol Med. 2024;173:108382.
28. Jaouedi N, Perales FJ, Buades JM, Boujnah N, Bouhlel MS. Prediction of human activities based on a New structure of Skeleton features and deep learning model. Sens (Basel). 2020;20(17):4944.
29. Liao Y, Vakanski A, Xian M, Paul D, Baker R. A review of computational approaches for evaluation of rehabilitation exercises. Comput Biol Med. 2020;119:103687.
30. Zhao Z, Yang T, Qin C, Zhao M, Zhao F, Li B, Liu J. Exploring the potential of the sit-to-stand test for self-assessment of physical condition in advanced knee osteoarthritis patients using computer vision. Front Public Health. 2024;12:1348236.
31. Ari Akgulys. A computerized recognition system for the home-based physiotherapy exercises using an RGBD camera. IEEE Trans Neural Syst Rehabilitation Engineering: Publication IEEE Eng Medicine& Biology Soc. 2014;22(6):1160–71.
32. Hoda M, Hoda Y, Hagea et al. Cloud-based rehabilitation and recovery prediction system for stroke patients. Cluster Comput. 2015;18(2):803–15.
33. Vakanskia, Jun HP, Pauld et al. A data set of human body movements for physical rehabilitation exercises. Data. 2018;3(1):2.
34. Liao Ylvakanskia, Xian M. A deep learning framework for assessing physical rehabilitation exercises. IEEE Trans Neural Syst Rehabilitation Eng. 2020;28(2):468–77.
35. Saiki Y, Kabata T, Ojima T, Kajino Y, Inoue D, Ohmori T, Yoshitani J, Ueno T, Yamamuro Y, Taninaka A, Kataoka T, Kubo N, Hayashi S, Tsuchiya H. Reliability and validity of OpenPose for measuring hip-knee-ankle angle in patients with knee osteoarthritis. Sci Rep. 2023;13(1):3297.
36. Chang YJ, Chen SF, Huang JD. A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. Res Dev Disabil. 2011;32(6):2566–70.

37. Li Y, Wu Y, Chen X, Chen H, Kong D, Tang H, Li S. Beyond Human detection: a Benchmark for detecting Common Human posture. Sens (Basel). 2023;23(19):8061.
38. Yao L, Xu H, Li A. Kinect-based rehabilitation exercises system: therapist involved approach. Biomed Mater Eng. 2014;24(6):2611–8.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.