International Scientific Technical and Economic Research (ISSN: 2959-1309)
Vol. 4, No. 1, (March 2026), pp. 188-199
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE DOI: 10.71451/ISTAER2609

Research on Automatic Evaluation Algorithm of Students' Sports Action Standardization Based on Computer Vision

Han Li[1], Yanan Sun[2]*

[1]*Sports Management College, Kyung Hee University, Yongin-si, Republic of Korea*

[2]*Sports Management College, Xi'an University of Architecture and Technology, Shaanxi, China*

Abstract: Aiming at the problems of strong subjectivity, lack of accuracy and difficulty in large-scale evaluation of students' sports action standardization, this paper proposes an automatic evaluation algorithm based on computer vision. First, a multi-perspective sports action dataset is constructed and an expert scoring system is designed; Secondly, key point sequences are extracted using an improved pose estimation model, and a multi-scale motion representation method is introduced to integrate joint-level, limb-level, and global features; Furthermore, a bias-aware alignment network is proposed to achieve adaptive modeling of spatiotemporal errors; Finally, a multi-task scoring model based on the fusion of GCN and Transformer is constructed to realize the normative classification and regression prediction of actions. The experimental results show that on the self-built data set, the MAE of this method is reduced to 0.318, which represents an improvement of approximately 29.6% over mainstream methods, the classification accuracy is 91.6%, and the correlation coefficient with expert score is 0.94. At the same time, in the cross-scenario test, the performance decreased by only 2.8%, which was significantly better than the comparison method. Ablation experiments and statistical tests validate the effectiveness of each module. The results show that this method has obvious advantages in accuracy, generalization ability and interpretability, and can provide technical support for intelligent physical education teaching and automatic evaluation.

Keywords: Computer vision; Sports movement evaluation; Attitude estimation; Multiscale feature; Graph convolution network; Action alignment; Deep learning

How to Cite: Li, H., & Sun, Y. (2026). Research on Automatic Evaluation Algorithm of Students' Sports Action Standardization Based on Computer Vision. *International Scientific Technical and Economic Research, 4*(1), 188–199. https://doi.org/10.71451/ISTAER2609

Article history: Received: 02 Jan 2026; **Revised:** 20 Feb 2026; **Accepted:** 18 Mar 2026; **Published:** 24 Mar 2026

1. INTRODUCTION

With the deepening of the concept of quality education, the role of physical education in students' comprehensive development has become increasingly prominent. The standardization of sports action not only directly affects sports performance, but also relates to students' health and sports safety [1],[2]. In the actual teaching process, standardized movements can effectively reduce the risk of sports injuries, improve the training efficiency, and promote students to form

***Corresponding author:** Yanan Sun, Sports Management College, Xi'an University of Architecture and Technology, Shaanxi, China. Email: sunyanan@xauat.edu.cn

good sports habits. Therefore, the scientific, objective and normative evaluation of students' sports movements has important educational significance and practical value [3]. In addition, in large-scale teaching scenarios, unified evaluation criteria help to improve teaching fairness and avoid evaluation bias caused by differences in teachers' experience. However, at present, the evaluation of movement standardization in physical education teaching mainly depends on manual observation and empirical judgment [4]. This method is not only subjective, but also difficult to quantify fine-grained movement differences, and it is difficult to meet the needs of intelligent and large-scale teaching.

The traditional manual evaluation method has obvious limitations. On the one hand, it is difficult for teachers to make comprehensive and accurate judgments on the complex movements of multiple students within a short time, and the evaluation efficiency is low; On the other hand, there are differences in the evaluation criteria among different teachers, leading to a lack of consistency in the results. In addition, for some actions with rapid dynamic changes or obvious minor deviations, manual evaluation makes it difficult to conduct continuous time-series analysis, which affects the accuracy of evaluation [5],[6]. With the development of information technology, computer vision technology has made significant progress in the field of human motion analysis. The pose estimation method based on deep learning can automatically extract human key points from video, and provide structured data basis for motion analysis; At the same time, the development of video understanding technology makes it possible to model complex actions in time and space [7],[8],[9]. These technologies provide new solutions for the automatic and objective evaluation of sports movements.

In recent years, scholars at home and abroad have carried out a large number of studies around human motion analysis. By extracting the key points of human skeleton and analyzing their spatial relationships, pose estimation-based methods enable structured representation of actions, offering the advantages of high computational efficiency and strong interpretability. However, such methods often rely on key point accuracy and are sensitive to complex scenes and occlusions. The video understanding method based on deep learning directly models the video sequence through convolutional neural network or transformer model, which can capture rich spatio-temporal information, and has achieved good results in the task of action recognition, but their application in evaluating action standardization remains relatively limited, and model complexity is high [10],[11],[12]. Aiming at the problem of sports movement evaluation, some studies attempt to combine template matching or score regression methods to evaluate movement quality, but the whole is still in the exploratory stage. Existing methods have several deficiencies: first, the lack of a unified and standardized evaluation index system for movement standardization makes horizontal comparisons across studies difficult; Secondly, the generalization ability of models is limited in cross-population and cross-scenario conditions, which is difficult to adapt to the complex changes in the actual teaching environment; Thirdly, they are not sufficiently sensitive to fine-grained deviations in movement (such as small changes in joint angle, inconsistent rhythm), so it is difficult to achieve high-precision evaluation.

In view of the above problems, this paper carries out a systematic study around the task of automatic evaluation of students' sports action standardization. First of all, the overall framework of action standardization evaluation for sports teaching scenes is constructed to realize the complete process from video input to score output; Secondly, a novel motion deviation modeling algorithm is proposed to describe the difference between the motion and the standard template across both temporal and spatial dimensions; On this basis, an evaluation model that integrates spatial structure and temporal dynamics is designed to improve the expression ability of complex actions; At the same time, a dataset containing multiple action types and expert-labeled scores was constructed, and a unified evaluation standard system was designed to enhance the standardization and reproducibility of the research; Finally, an interpretable evaluation mechanism is introduced to enable localization and visual analysis of key erroneous actions, so as to provide intuitive and operable feedback for physical education teaching. Through the above research, this paper aims to provide a high-precision, strong

generalization and interpretable technical scheme for intelligent physical education teaching and automatic evaluation.

2. DATA CONSTRUCTION & PREPROCESSING

In this study, the construction and preprocessing of the dataset form the foundation for achieving accurate automatic evaluation of sports movement standardization. Due to the diversity and complexity of sports movements and the individual differences of human movement, the quality of data construction directly affects the performance and generalization capability of subsequent models. This section describes in detail the data set construction method, data annotation strategy, data preprocessing and data set division method, in order to provide standardized and high-quality input data for subsequent model training.

2.1. Dataset construction

In order to ensure the representativeness and diversity of the data set, this study focuses on four common sports movements. These movements cover essential physical education content and hold significant value for both teaching and evaluation. Specifically, it includes: running, which is used to evaluate the running posture and gait coordination of students; The long jump is used to evaluate the standardization of take-off and landing movements, particularly the angle and force control during take-off; Sit-ups were used to evaluate core muscle group engagement and movement accuracy; Push up is used to evaluate upper limb strength and movement stability. These four kinds of movements involve different motor skills, which can represent the movement standardization of students in various types of sports activities, so they constitute the core of the data set of this study.

In the process of data collection, the research team used multiple camera perspectives to synchronously collect data from the front, rear, left, right, and oblique directions to ensure more comprehensive and multi-dimensional motion information. The video of each action contains data from at least 2 to 3 different perspectives. This multi perspective setting provides important support for the subsequent capture of joint position deviation from different angles.

In addition to the perspective differences, the dataset also covers variations across different individuals as comprehensively as possible, including factors such as age, height, and gender. Students of different body types may show different movement patterns when performing the same action. Therefore, in the process of data set construction, special attention is paid to including students with diverse body types and physiological characteristics, so as to ensure that the constructed algorithm has better universality and robustness in practical application.

In order to provide high-quality annotation data for the subsequent evaluation model, this study invited several physical education teachers and sports training experts to score each action according to standard action specifications. The scoring criteria are mainly divided into the following three categories: action accuracy, that is, whether the action conforms to the predetermined standard action trajectory and whether there is obvious deviation; Movement fluency, i.e., whether the movement is coherent and smooth without unnatural pauses or abrupt changes; The range of motion, that is, whether the range of motion meets the requirements, especially in the long jump, sit ups and other movements requiring a specific range, the standard range is very important for the evaluation results.

A five-point grading system is adopted for scoring, in which a score of 1 indicates highly irregular, while a score of 5 indicates fully standardized. The scoring process is completed by the combination of video playback and motion trajectory annotation to ensure that the scoring results are based on both intuitive observation and accurate motion data analysis. **Table 1** shows the specific scoring of different action types under each scoring standard.

Table 1. Score of movement standardization

Action type	Action accuracy score	Movement fluency score	Range of motion score
Run	4	5	4
Long jump	5	4	5
Sit ups	4	4	4
Push ups	3	4	4

It can be observed that different action types exhibit distinct characteristics across the three dimensions of action accuracy, movement fluency, and range of motion. Running achieves a full score of 5 in movement fluency, indicating that experts generally consider rhythmic consistency and limb coordination as the most critical aspects of running movements, while the scores for action accuracy and range of motion are both 4, suggesting that some individual variations exist in posture details and stride control. Long jump receives the highest score of 5 in both action accuracy and range of motion, reflecting that this action imposes the strictest requirements on take-off angle, aerial posture, and landing trajectory, with the adequacy of range of motion serving as a key indicator of standardization; the fluency score of 4 indicates that there is still room for improvement in the smoothness of action transitions for some samples. Sit-ups achieve a score of 4 across all three dimensions, demonstrating the most balanced performance; this action has a relatively simple structure and strong repetitiveness, with consistent scores across dimensions, indicating a high degree of standardization and minimal variation among evaluation criteria. Push-ups receive a score of 3 in action accuracy, making it the only action with a score below 4 among the four types, suggesting that experts commonly identify postural deviations during execution (such as sagging waist or elbow flaring), making it a more challenging action type for standardization assessment; the scores for movement fluency and range of motion are both 4, reflecting relatively stable performance in these aspects.

2.2 Data annotation strategy

In this study, we extract key points based on human pose estimation algorithms in computer vision (such as OpenPose or HRNet) and annotate them. Key points mainly include the main joints of the human body (such as shoulders, elbows, kneSes, ankles, etc.), and the position of each joint is represented by two-dimensional or three-dimensional coordinates.

For each kind of sports action, the position of key points in the action is calculated to determine whether the action conforms to the standard trajectory [13]. Key point deviation is one of the key factors in evaluating action standardization in this study. The deviation is expressed as:

$$\text{Deviation} = \sqrt{\sum_{i=1}^{n} (x_i - x_i^*)^2 + (y_i - y_i^*)^2} \quad (1)$$

Where, x_i, y_i represents the actual coordinates of the ith key point, x_i^*, y_i^* represents the ideal coordinates in the standard action, and n is the number of key points.

Each sports action is divided into several stages, and performance at different stages is annotated. For example, the long jump consists of four stages: approach, take-off, flight, and landing, while sit-ups can be divided into three stages: rise, hold, and descent. The movement standardization of each stage was scored independently, and the comprehensive evaluation was carried out in combination with the overall movement performance. Stage-level annotation

helps better quantify the details of each action and improves evaluation accuracy.

Based on the annotation of experts, we designed a standardized scoring label. Each action receives a comprehensive score based on stage-level and overall evaluations. The calculation formula of the comprehensive score is:

$$S_{\text{total}} = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{2}$$

Among them, S_i is the action standardization score of stage i, n is the number of action stages, and S_{total} is the overall action score.

2.3 Data preprocessing

Video data needs frame sampling to reduce redundant information and improve processing efficiency. In this study, a temporal frame sampling method is adopted, selecting 10 frames per second for processing, which can achieve an effective balance between the amount of calculation and accuracy requirements. In the sampling process, the time interval between frames is 0.1 seconds, ensuring that motion changes are fully captured while avoiding the loss of key motion information due to a low sampling rate.

In order to improve the generalization ability of the model and avoid over fitting, this study performed various data augmentation techniques. Specific enhancement techniques include: rotation, that is, random rotation of the video frame to simulate the action performance under different angles; Zooming, such as randomly zooming images to simulate action performance at different distances; Occlusion, that is to simulate the occlusion phenomenon in the actual situation, and block some areas artificially, so as to enhance the robustness of the model in the case of partial information loss. Through the above enhancement means, the diversity of training data is effectively expanded, and the adaptability of the model in complex scenes is improved.

Due to the body shape differences of different individuals, the coordinates of posture key points are quite different among different students [14],[15]. In order to eliminate the influence of body shape on model training, the coordinates of all key points were normalized and aligned to a standardized coordinate system. The specific formula is:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, y' = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \tag{3}$$

Where x', y' are the original coordinates, andx', y' are the normalized coordinates, $x_{\min}, x_{\max}$ are the minimum and maximum values of x coordinates in the dataset, and other variables are the same.

2.4. Dataset partitioning

In order to train, validate and test the model, we divide the dataset into training, validation, and test sets. Generally, the training set accounts for 70% of the total data set, the verification set accounts for 15%, and the test set accounts for 15%. In addition, in order to enhance the generalization ability of the model, we also conducted a cross-scenario test design, that is, the test set includes data from different scenarios to verify model performance across various environments. Table 2 shows the data set division.

Table 2. Dataset partitioning

Dataset type	Quantity (video samples)	Proportion
Training set	3500	70%
Validation set	750	15%
Test set	750	15%

Through this data partitioning strategy, we ensure that the model can be generalized in different scenarios and different individuals, and also ensure the diversity of the data set.

To sum up, this study has laid a solid foundation for the follow-up model training through multi perspective, multi human difference data acquisition, accurate annotation strategy and reasonable data preprocessing methods. Through data enhancement, key point normalization and other technologies, the robustness and generalization ability of the model are improved. In addition, the data set partitioning strategy also ensures the objectivity and comprehensiveness of the evaluation model, and provides a solid experimental basis for subsequent research.

3. CORE METHODOLOGY

Focusing on the core issue of "automatic evaluation of students' sports action standardization", this study constructed a complete method system from posture estimation, time series modeling, deviation analysis and score prediction. The overall method is based on human key point sequences, extracts multi-scale spatiotemporal features using deep learning models, and achieves refined deviation modeling in combination with standard action templates, ultimately enabling high-precision and interpretable action standardization evaluation.

In terms of pose estimation and time series modeling, an improved High-Resolution Network (HRNet) is used as the base skeleton extraction model [16],[17]. Let the input video frame be I_t, and the set of human key points output by the pose estimation network is:

$$P_t = \{(x_i^t, y_i^t, c_i^t)\}_{i=1}^{N} \tag{4}$$

Where N is the number of key points, (x_i^t, y_i^t) is the spatial coordinates of the ith key point at time t, and c_i^t is the confidence level. To address the lack of robustness of traditional HRNet to occlusions and fast motions in complex sports scenes, this paper introduces a cross-scale feature fusion module and enhances the response of key joint regions using an attention mechanism.

$$F = \sum_{s=1}^{S} \alpha_s \cdot F_s \tag{5}$$

Where F_s is the feature map of the s-th scale, and α_s is the learnable weight.

In time series modeling, the sequence of key points in consecutive frames is represented as:

$$X = [P_1, P_2, \dots, P_T] \in \mathbb{R}^{T \times N \times 2} \tag{6}$$

Time dependencies are modeled using LSTM or Transformer. Taking Transformer as an example, its self attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{7}$$

Where Q, K, V are the query, key and value matrices respectively, and d is the characteristic dimension. In addition, the action sequence is divided into K stages by introducing the action stage division mechanism.

$$X = \bigcup_{k=1}^{K} X^{(k)} \tag{8}$$

Each stage is modeled by an independent encoder to improve the ability to capture fine-grained changes in motion.

For action standardization feature representation, this paper proposes a multi-scale motion representation method. Firstly, the joint angle feature is defined. For three adjacent joints i, j, k, The angle is defined as:

$$\theta_{ijk} = \arccos\left(\frac{(p_i - p_j) \cdot (p_k - p_j)}{\| p_i - p_j \| \| p_k - p_j \|}\right) \tag{9}$$

Where p_i represents the ith joint coordinate. The bone topology is represented by the graph structure $G = (V, E)$, where V is the joint node set and E is the bone connection relationship. Temporal dynamic characteristics are described by velocity and acceleration [18]:

$$v_i^t = p_i^t - p_i^{t-1}, a_i^t = v_i^t - v_i^{t-1} \tag{10}$$

The consistency of movement rhythm is further defined as:

$$R = \frac{1}{T-1} \sum_{t=2}^{T} \| v^t - \bar{v} \| \tag{11}$$

Where $\bar{v}$ is the average velocity. Multiscale representation divides features into joint level (single point), limb level (bone pair) and global level (global motion) [19].

$$F_{multi} = \lambda_1 F_{joint} + \lambda_2 F_{limb} + \lambda_3 F_{global} \tag{12}$$

Where λ_i is the weight parameter.

For motion deviation modeling, the standard motion template is constructed first. Set the standard action sequence as:

$$X^* = [P_1^*, P_2^*, \ldots, P_T^*] \tag{13}$$

Derived from expert annotations or statistical averaging:

$$P_t^* = \frac{1}{M} \sum_{m=1}^{M} P_t^{(m)} \tag{14}$$

Where M is the number of samples. To address temporal misalignment, an improved Dynamic Time Warping (DTW) method is introduced [20],[21],[22]:

$$DTW(X, X^*) = \min_{\pi} \sum_{(t,s)\in\pi} \| P_t - P_s^* \| \tag{15}$$

Where π is the alignment path. An attention mechanism is further incorporated to construct alignment weights:

$$A_{ts} = \frac{\exp(-\| P_t - P_s^* \|)}{\sum_s \exp(-\| P_t - P_s^* \|)} \tag{16}$$

On this basis, a bias-aware alignment network is proposed, whose overall bias is defined

as:

$$D = \sum_{t=1}^{T} \sum_{s=1}^{T} A_{ts} \cdot (\alpha \parallel P_t - P_s^* \parallel + \beta \parallel v_t - v_s^* \parallel) \tag{17}$$

Where α, β are space and time weights to achieve adaptive error fusion.

For the standardization scoring model, a multi-task learning framework is constructed, and the classification and regression tasks are carried out at the same time. The model outputs are defined as:

$$\hat{y}_{cls} = \sigma(W_c F + b_c), \hat{y}_{reg} = W_r F + b_r \tag{18}$$

Where σ denotes the sigmoid function. The skeleton structure is modeled using a graph convolutional network, and its propagation rules are:

$$H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}\right) \tag{19}$$

Where A is the adjacency matrix and D is the degree matrix. The Transformer further captures long-term dependencies. A hierarchical scoring mechanism is proposed, with the final score defined as:

$$S = \gamma_1 S_{joint} + \gamma_2 S_{stage} + \gamma_3 S_{global} \tag{20}$$

S_{joint} is the joint score, S_{stage} is the stage score, and S_{global} is the overall score.

Finally, in the interpretability evaluation mechanism, key erroneous regions are identified using gradient-based response:

$$M_i = \frac{\partial S}{\partial p_i} \tag{21}$$

Used to locate key wrong joints. Further construct the deviation thermodynamic diagram [23]:

$$H_t = \sum_{i=1}^{N} w_i \cdot \parallel p_i^t - p_i^{t*} \parallel \tag{22}$$

Where w_i is the joint importance weight. Based on the deviation distribution, a teaching feedback function is generated:

$$Feedback = f(H, \theta, v) \tag{23}$$

Enabling automatic diagnosis and natural language feedback on students' movement issues.

To sum up, this method achieves high-precision and interpretable automatic evaluation of students' sports movement standardization through the complete technical chain of "posture estimation-multi-scale feature modeling-deviation alignment-hierarchical scoring-interpretable feedback", demonstrating clear innovation in both model architecture and algorithmic design.

4.EXPERIMENTS & RESULTS

In order to comprehensively verify the effectiveness and advancement of the proposed method in the task of automatic assessment of students' sports action standardization, this study systematically designed the experimental setup, baseline methods, evaluation metrics, and other

aspects, and analyzed the performance of the model through multi-dimensional experiments (main experiment, generalization experiment, ablation experiment and statistical test).

In terms of experimental setup, all experiments were completed on a unified platform. The hardware environment includes an NVIDIA RTX 4090 GPU (24GB video memory), an Intel Xeon Gold 6230 CPU, and 128GB of memory; the software environment is based on Ubuntu 22.04 and accelerated using PyTorch 2.1 and CUDA 12.1. The model training batch size is set to 32, and the initial learning rate is 1×10^{-4}. The Adam optimizer is used, and its update rule is [24],[25]:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{24}$$

Where θ_t is the parameter, η is the learning rate, m_t, v_t are the first-order and second-order moment estimates, respectively.

The compared methods include traditional approaches and deep learning methods. Traditional methods include DTW-based template matching (DTW-Match) and an SVM regression model using handcrafted features (angle + velocity) (Handcraft+SVM); Deep learning methods include ST-GCN, PoseC3D, and a Transformer-based video understanding model. evaluation metrics are defined for both regression and classification tasks. The mean absolute error (MAE) is defined as [26],[27]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i | \tag{25}$$

Root mean square error (RMSE) is [28]:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{26}$$

The classification accuracy is [29]:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

F1-score is defined as [30]:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{28}$$

In addition, to measure the consistency between the model and expert scores, Pearson correlation coefficient is introduced [31]:

$$r = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}} \tag{29}$$

The overall performance of different methods in the main experiment is shown in Table 3. It can be observed that the proposed method significantly outperforms the compared methods across all metrics.

Table 3. Comparison of main experimental results

Method	MAE ↓	RMSE ↓	Acc ↑	F1 ↑	Correlation coefficient r ↑
DTW-Match	0.842	1.103	71.2%	0.702	0.68
Handcraft+SVM	0.735	0.964	75.6%	0.741	0.72
ST-GCN	0.512	0.703	84.3%	0.831	0.86
PoseC3D	0.476	0.661	85.9%	0.847	0.88
Transformer	0.452	0.638	87.2%	0.861	0.90
Method in this paper	0.318	0.472	91.6%	0.907	0.94

As shown in Table 3, the MAE of the proposed method is reduced by approximately 29.6%, indicating a substantial reduction in scoring error; The correlation coefficient reached 0.94, indicating that it was highly consistent with the expert score.

As shown in Table 4, the proposed method performs stably across six different movement types, with overall error kept at a low level.

Table 4. Performance analysis of different action categories

Action type	MAE	Acc	F1	r
Running	0.302	92.1%	0.91	0.95
Long jump	0.336	90.8%	0.90	0.93
Sit ups	0.321	91.5%	0.91	0.94
Push ups	0.349	90.2%	0.89	0.92
Squat	0.315	91.8%	0.91	0.94
Pull up	0.362	89.7%	0.88	0.91

Running has the lowest MAE (0.302) and the highest accuracy (92.1%), indicating high consistency in postural features and the model's strongest ability to recognize its standardization. Pull-ups, with their complex structure and high joint freedom, have a slightly higher MAE (0.362) and relatively lower accuracy (89.7%), reflecting the continued challenges in standardized modeling of this type of movement. Overall, the model's F1 score is no lower than 0.88 across all movement types, and the correlation coefficient (r) is higher than 0.91, indicating a high degree of consistency between model scores and expert scores.

In the generalization ability test, cross-scenario data (indoor→outdoor) is introduced, and the performance variation trend is shown in Figure 1.

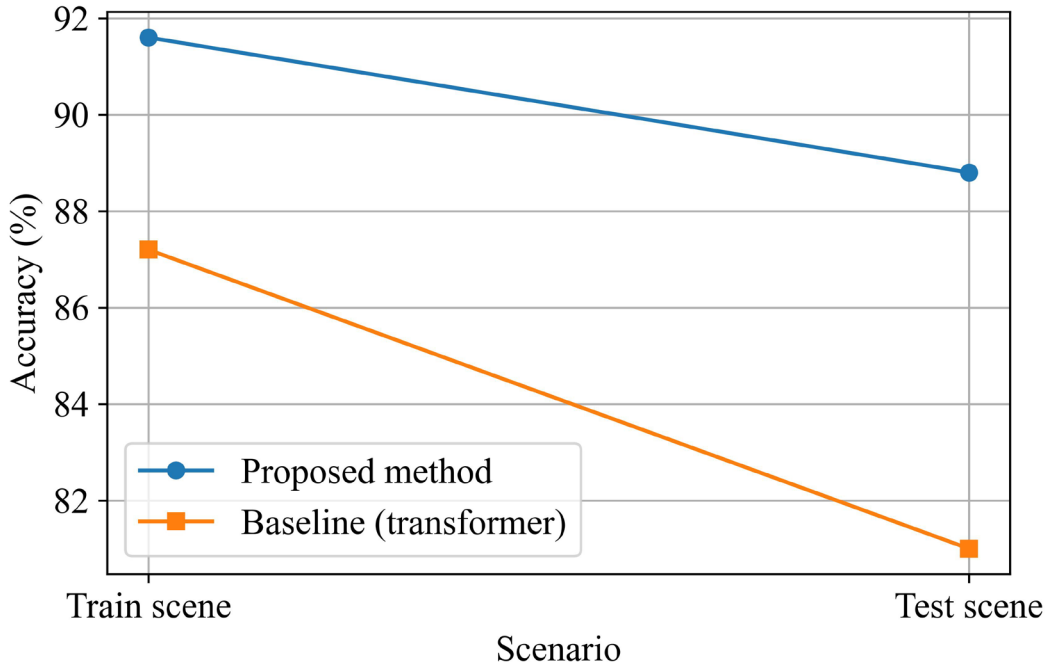

Figure 1. Variation curve of generalization performance

The results show that the performance of the proposed method drops by only about 2.8%, significantly outperforming the comparison models (which exhibit an average drop of over 6%), demonstrating strong cross-scenario generalization ability.

In the ablation experiment, to verify the contribution of each module, key components were gradually removed, and the results are presented in Table 5.

Table 5. Ablation results

Model variants	MAE	Acc
Complete model	0.318	91.6%
Remove multiscale representation	0.401	88.7%
Remove deviation alignment network	0.437	87.9%
Remove hierarchical scoring mechanism	0.389	89.2%
Use normal DTW instead	0.462	86.5%
Use GCN only	0.428	87.1%

It can be seen that the multi-scale representation and deviation alignment modules contribute the most, each reducing error by over 20%. Further analysis of the impact of feature representation methods is shown in Figure 2:

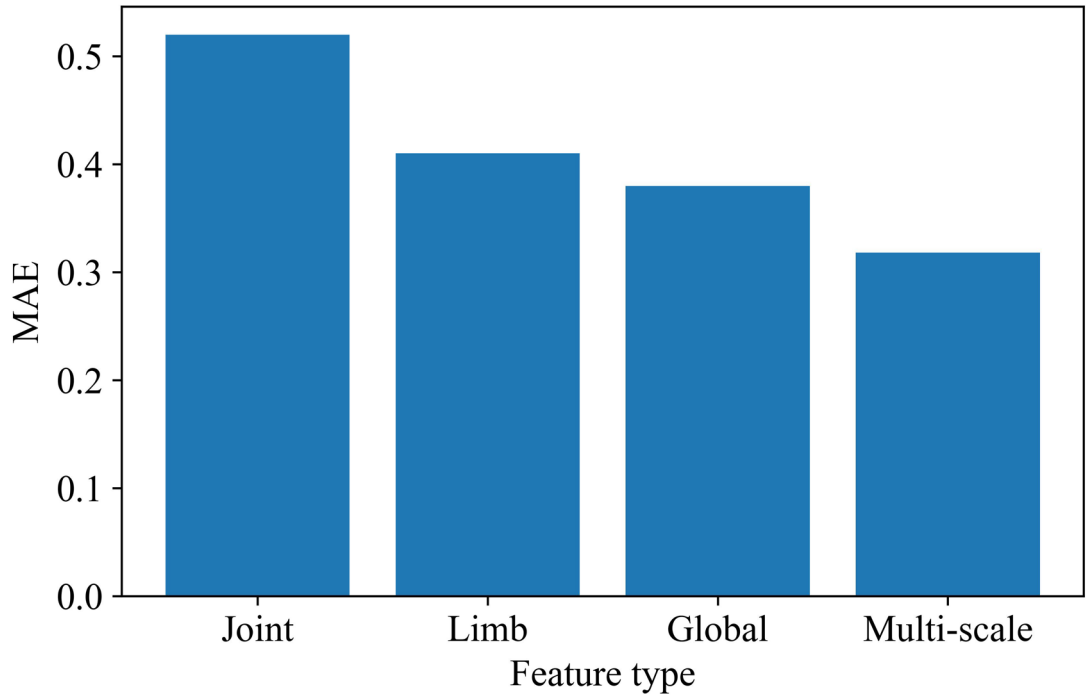

Figure 2. Comparison of feature representation (MAE)

It shows that multi-scale fusion significantly outperforms single-feature representations.

In the analysis of the alignment algorithm, the error function after attention-based alignment is introduced [32],[33]:

$$L_{align} = \sum_{t,s} A_{ts} \parallel P_t - P_s^* \parallel^2 \tag{30}$$

Experiments show that the error is reduced by approximately 18% compared with traditional DTW.

In the model structure analysis, as shown in Figure 3, the GCN + Transformer architecture is clearly superior to single-structure alternatives:

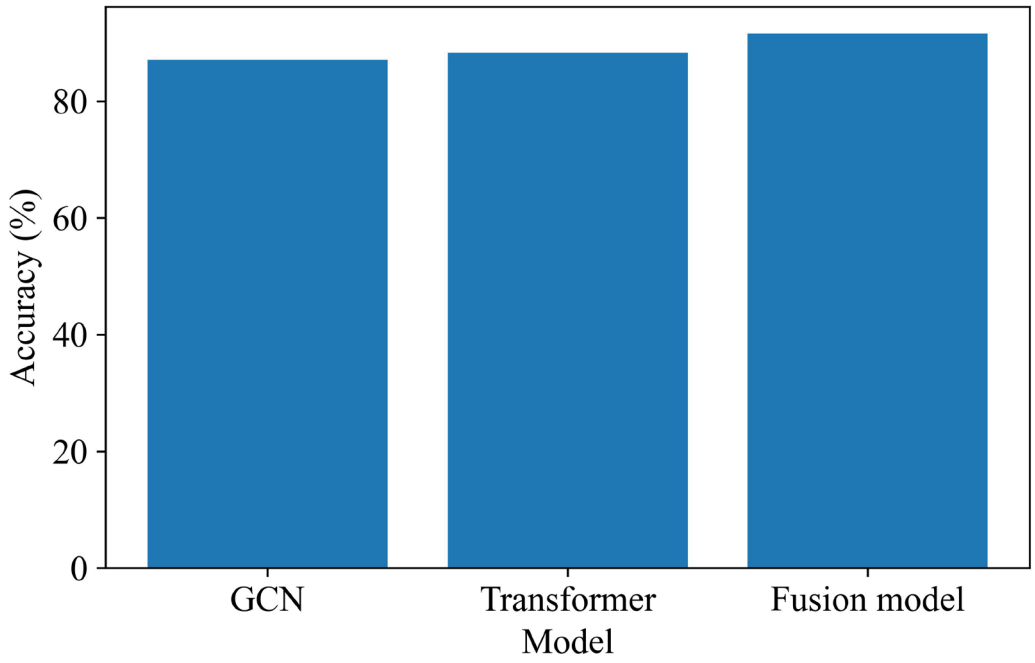

Figure 3. Model structure comparison (Acc)

In the visual analysis, key erroneous areas are displayed using deviation heatmaps, as shown in Figure 4:

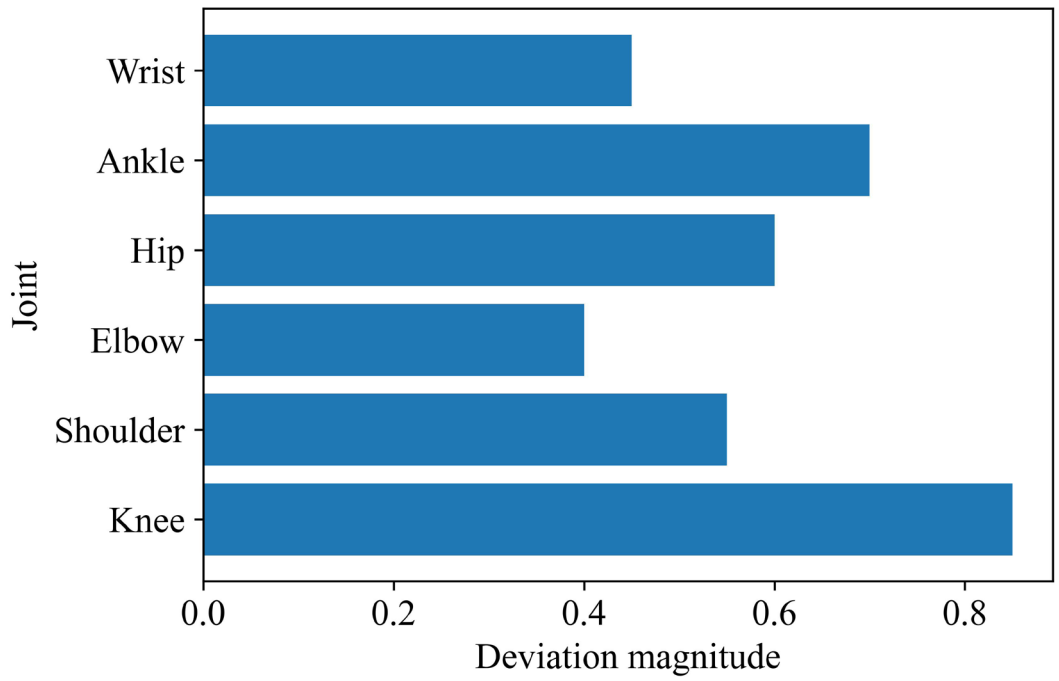

Figure 4. Thermal diagram of action deviation

The heatmap is computed as $H_i = \| p_i - p_i^* \|$, which directly reflects that errors are concentrated in the knee joint area. Error case analysis shows that the model still exhibits misjudgments when motion rhythm is abnormal or occlusion is severe. Through gradient-based

interpretation, $M_i = \frac{\partial S}{\partial p_i}$ can locate the model's regions of interest, thereby improving interpretability.

Finally, to verify the statistical significance of the performance improvement, a paired t-test was used [34],[35].

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{31}$$

Where $\bar{d}$ is the mean value of the difference and s_d is the standard deviation. The experimental results are shown in Table 6.

Table 6. Significance test results

Comparison method	P-value
vs ST-GCN	0.003
vs PoseC3D	0.001
vs Transformer	0.0007
vs SVM	<0.0001
vs DTW	<0.0001

All p-values are less than 0.01, indicating that the performance improvement is statistically significant. Additionally, the Wilcoxon test was used to verify the reliability of the model's advantages.

To sum up, the experimental results fully demonstrate the effectiveness and superiority of the proposed method in terms of accuracy, generalization ability, module contributions, and statistical significance, especially in fine-grained motion evaluation and expert consistency, demonstrating strong potential for practical application.

5. DISCUSSION

Based on the overall experimental results and methodological design, the computer vision-based automatic evaluation method for students' sports movement standardization proposed in this paper demonstrates clear advantages across several key dimensions. Firstly, in terms of evaluation accuracy, by integrating multi-scale motion representation and a deviation-aware alignment mechanism, the model can effectively capture fine-grained motion differences and achieve accurate modeling of joint angles, motion trajectories, and rhythmic variations. Compared with traditional methods based on template matching or single features, this model maintains stable prediction performance in complex action scenes while significantly reducing scoring errors. At the same time, the multi-task learning framework jointly optimizes classification and regression tasks, enabling the model to strike a good balance between categorical criteria and quantitative scoring, thereby improving overall evaluation accuracy. In terms of generalization ability, thanks to the multi-perspective data construction and spatiotemporal feature fusion strategy, the model can adapt to different body structures, movement styles, and environmental changes, showing strong robustness in cross-scenario testing. This generalization ability is of great significance to the diverse student groups in the actual teaching environment. In addition, this method also offers outstanding advantages in interpretability. By incorporating deviation visualization and key joint localization mechanisms,

the model can not only provide scoring results but also clearly indicate the specific areas and reasons for non-standard movements, offering intuitive feedback for teachers and students and enhancing the system's practicality and reliability.

Although this method has achieved good results in many aspects, it still has some limitations. First of all, the model is highly dependent on data quality, particularly in the pose estimation stage, if key point detection contains errors or omissions, it directly affects subsequent feature extraction and scoring results. In practical applications, complex lighting, occlusions, or background interference may still reduce system stability. Secondly, the issue of multi-person interference is common in real-world physical education settings. For example, in the collective training or classroom environment, multi-person interactions may lead to key point confusion or tracking errors, thereby affecting assessment accuracy. the current method mainly focuses on modeling single-person actions, and processing multi-person scenes still requires further research. In addition, in terms of real-time performance, although the proposed model performs well in offline evaluation, its computational complexity remains high due to the integration of multiple deep network modules, posing challenges in resource-constrained devices or real-time feedback scenarios. Therefore, how to reduce model complexity and improve inference efficiency while maintaining accuracy is a key issue to be addressed in future work.

From the perspective of practical application, this study holds significant value for deployment. In the field of physical education, this method can assist teachers in achieving objective evaluation of students' movements, reducing teaching burden, and improving teaching efficiency and fairness. through automatic scoring and visual feedback, students can promptly identify deficiencies in their movements and make targeted improvements. In the aspect of intelligent evaluation system, this method can be integrated as a core module into smart campus or online physical education platforms to enable large-scale, standardized action evaluation and score management, promoting the digital transformation of physical education. At the same time, the technology also holds potential application value in health monitoring and exercise guidance. For example, in daily fitness or rehabilitation training, the system can perform real-time analysis of user movements and provide normative feedback, thereby reducing the risk of sports injuries and enhancing training effectiveness. In general, this method not only enriches the theoretical framework of motion evaluation modeling but also provides a feasible technical pathway for intelligent sports and health management in practice.

6. CONCLUSION

This paper addresses the key problem of automatic evaluation of students' sports movement standardization and proposes a systematic computer vision-based solution. By constructing a complete technical framework encompassing pose estimation, time series modeling, deviation analysis, and score prediction, automatic and refined evaluation of sports movement standardization is achieved. At the methodological level, this paper designs a multi-scale feature representation that integrates spatial structure and temporal dynamics, effectively enhancing the model's representational capacity for complex actions; At the same time, by introducing a deviation-aware alignment mechanism, precise modeling of spatiotemporal discrepancies in actions is achieved, improving both the sensitivity and stability of the evaluation. In addition, combined with a graph convolutional network and time series modeling architecture, a unified scoring model is constructed to achieve collaborative optimization between action recognition and score prediction. The experimental results show that this method is superior to the existing mainstream methods in many evaluation indexes, which not only significantly reduces scoring error but also achieves a high degree of consistency with expert scores, which verifies the effectiveness and practicability of the model. At the same time, the stable performance in the cross -scenario testing further demonstrates the method's strong generalization capability.

Although this study has made certain progress, there remains room for further expansion in more complex and realistic application environments. Future research can explore multimodal fusion, combining video data with inertial measurement unit (IMU) and other sensor data to compensate for the limitations of visual-only information in occluded or complex environments, thereby improving system robustness and accuracy In addition, given the high cost of obtaining annotated data in physical education, methods such as few-shot learning and self-supervised learning can be explored to reduce dependence on large-scale labeled datasets and improve model adaptability to new action categories. At the system application level, real-time online evaluation remains an important development direction. In the future, low-latency, high-efficiency real-time feedback can be achieved through model lightweighting, edge computing, and other techniques to meet the immediate needs of classroom instruction or training scenarios. At the same time, cross-scenario transfer learning is also a direction worthy of in-depth investigation. By incorporating domain adaptation or transfer learning strategies, the model can rapidly adapt to different environments and individuals, further enhancing its practical application value.

In general, the research presented in this paper provides a technical pathway with high accuracy and good scalability for the automatic evaluation of sports movements. In the future, with the development of multimodal perception, lightweight models, and intelligent education platforms, such methods are expected to play an increasingly important role in intelligent sports, health monitoring, and personalized training.

Abbreviations

HRNet, High-Resolution Network;
LSTM, Long Short-Term Memory;
GCN, Graph Convolutional Network;
DTW, Dynamic Time Warping;
SVM, Support Vector Machine;
ST-GCN, Spatial Temporal Graph Convolutional Network;
MAE, Mean Absolute Error;
RMSE, Root Mean Square Error;
Acc, Accuracy;
F1, F1-Score;
TP, True Positive;
TN, True Negative;
FP, False Positive;
FN, False Negative;
GPU, Graphics Processing Unit;
CUDA, Compute Unified Device Architecture;
Adam, Adaptive Moment Estimation;
IMU, Inertial Measurement Unit.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **H.L.**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision. **Y.S.**: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **Y.S.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Chang, C. J., Putukian, M., Aerni, G., Diamond, A. B., Hong, E. S., Ingram, Y. M., ... & Wolanin, A. T. (2020). Mental health issues and psychological factors in athletes: detection, management, effect on performance, and prevention: American medical society for sports medicine position statement. *Clinical Journal of Sport Medicine*, *30*(2), e61-e87. DOI: **https://doi.org/10.1136/bjsports-2019-101583**

[2] Jo, K. H., Lee, S. M., So, W. Y., & Lee, E. J. (2023, June). Mediating effect of sports safety awareness between sports activity habits and the intention to complete safety education among Korean adolescents. In *Healthcare* (Vol. 11, No. 13, p. 1891). MDPI. DOI: **https://doi.org/10.3390/healthcare11131891**

[3] Potop, V., Manolachi, V., Mihailescu, L. E., Manolachi, V., & Kulbayev, A. (2022). Knowledge of the fundamentals necessary for the scientific research activity in the field of Physical Education and Sports Science. *Journal of Physical Education and Sport*, *22*(8), 1922-1926. DOI: **https://doi.org/10.7752/jpes.2022.08243**

[4] Hsia, L. H., Hwang, G. J., Lin, Y. N., & Hwang, J. P. (2025). Artificial intelligence-supported physical education during the pandemic: a physical skill auto-assessment and feedback approach based on a reflection-promoting mechanism: L.-H. Hsia et al. *Educational technology research and development*, *73*(3), 1429-1450. DOI: **https://doi.org/10.1007/s11423-025-10452-7**

[5] Usmani, U. A., Aziz, I. A., Jaafar, J., & Watada, J. (2024). Deep learning for anomaly detection in time-series data: An analysis of techniques, review of applications, and guidelines for future research. *IEEE Access*, *12*, 174564-174590. DOI: **https://doi.org/10.1109/ACCESS.2024.3495819**

[6] Middlehurst, M., Schäfer, P., & Bagnall, A. (2024). Bake off redux: a review and experimental evaluation of recent time series classification algorithms: M. Middlehurst et al. *Data Mining and Knowledge Discovery*, *38*(4), 1958-2031. DOI: **https://doi.org/10.1007/s10618-024-01022-1**

[7] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., ... & Shah, M. (2023). Deep learning-based human pose estimation: A survey. *ACM computing surveys*, *56*(1), 1-37. DOI: **https://doi.org/10.1145/3603618**

[8] Lan, G., Wu, Y., Hu, F., & Hao, Q. (2022). Vision-based human pose estimation via deep learning: A survey. *IEEE Transactions on Human-Machine Systems*, *53*(1), 253-268. DOI: **https://doi.org/10.1109/TPAMI.2008.106**

[9] Zhang, X., Zhou, Z., Han, Y., Meng, H., Yang, M., & Rajasegarar, S. (2023). Deep learning-based real-time 3D human pose estimation. *Engineering Applications of Artificial Intelligence*, *119*, 105813. DOI: **https://doi.org/10.5281/zenodo.17880888**

[10] Hussain, A., Hussain, T., Ullah, W., & Baik, S. W. (2022). Vision transformer and deep sequence learning for human activity recognition in surveillance videos. *Computational Intelligence and Neuroscience*, *2022*(1), 3454167. DOI:

https://doi.org/10.1155/2022/3454167

[11] Li, J., Liu, X., Zhang, W., Zhang, M., Song, J., & Sebe, N. (2020). Spatio-temporal attention networks for action recognition and detection. *IEEE Transactions on Multimedia*, *22*(11), 2990-3001. DOI: **https://doi.org/10.1109/TMM.2020.2965434**

[12] Huang, Y., Zhao, H., Zhou, Y., Riedel, T., & Beigl, M. (2023, November). Standardizing Your Training Process for Human Activity Recognition Models–A Comprehensive Review in the Tunable Factors. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services* (pp. 15-27). Cham: Springer Nature Switzerland. DOI: **https://doi.org/10.1007/978-3-031-63992-0_2**

[13] Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, *54*(3), 2259-2322. DOI: **https://doi.org/10.1007/s10462-020-09904-8**

[14] Ohlendorf, D., Avaniadi, I., Adjami, F., Christian, W., Doerry, C., Fay, V., ... & Maurer-Grubinger, C. (2023). Standard values of the upper body posture in healthy adults with special regard to age, sex and BMI. *Scientific Reports*, *13*(1), 873. DOI: **https://doi.org/10.1038/s41598-023-27976-8**

[15] Tao, W., Du, B., Li, B., He, W., & Sun, H. J. (2020). Body-posture recognition by undergraduate students majoring in physical education and other disciplines. *Frontiers in psychology*, *11*, 505543. DOI: **https://doi.org/10.3389/fpsyg.2020.505543**

[16] Liu, Y., Zhou, G., He, W., Zhu, H., & Cui, Y. (2025). DE-HRNet: Detail enhanced high-resolution network for human pose estimation. *PLoS One*, *20*(9), e0325540. DOI: **https://doi.org/10.1371/journal.pone.0325540**

[17] Wang, Y., Wang, R., Shi, H., & Liu, D. (2024). MS-HRNet: multi-scale high-resolution network for human pose estimation: Y. Wang et al. *The Journal of Supercomputing*, *80*(12), 17269-17291. DOI: **https://doi.org/10.1007/s11227-024-06125-6**

[18] Guo, X., Li, C., Luo, Z., & Cao, D. (2024). Identification of track irregularities with the multi-sensor acceleration measurements of vehicle dynamic responses. *Vehicle System Dynamics*, *62*(4), 906-931. DOI: **https://doi.org/10.1080/00423114.2023.2200193**

[19] Zhu, J., Zhang, Z., Liu, R., Ren, M., & Ma, G. (2025). Multiscale Modeling and Reconstruction of Joint Motion: Finite Element Optimization Based on Particle Swarm Algorithm. *IEEE Access*. DOI: **https://doi.org/10.1109/ACCESS.2025.3553469**

[20] Yang, D., Shaw, T., & Tsai, T. J. (2022). A study of parallelizable alternatives to dynamic time warping for aligning long sequences. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2117-2127. DOI: **https://doi.org/10.1109/TASLP.2022.3180673**

[21] Liu, Y., Zhang, Y. A., Zeng, M., & Zhao, J. (2024). A novel distance measure based on dynamic time warping to improve time series classification. *Information Sciences*, *656*, 119921. DOI: **https://doi.org/10.1016/j.ins.2023.119921**

[22] Kraprayoon, J., Pham, A., & Tsai, T. J. (2024). Improving the robustness of DTW to global time warping conditions in audio synchronization. *Applied Sciences*, *14*(4), 1459. DOI: **https://doi.org/10.3390/app14041459**

[23] Laughlin, D. E., & Massalski, T. B. (2021). Construction of equilibrium phase diagrams: Some errors to be avoided. *Progress in Materials Science*, *120*, 100715. DOI: **https://doi.org/10.1016/j.pmatsci.2020.100715**

[24] Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications*, *35*(23), 17095-17112. DOI: **https://doi.org/10.1007/s00521-023-08568-z**

[25] Yi, D., Ahn, J., & Ji, S. (2020). An effective optimization method for machine learning based on ADAM. *Applied Sciences*, *10*(3), 1073.DOI: **https://doi.org/10.3390/app10031073**

[26] Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PloS one*, *18*(2), e0279774. DOI: **https://doi.org/10.1371/journal.pone.0279774**

[27] Warneke, K., Siegel, S. D., Afonso, J., & Wallot, S. (2025). What the mean absolute percentage error (MAPE) should adopt from Bland–Altman analyses. *German Journal of Exercise and Sport Research*, 1-8. DOI: **https://doi.org/10.1007/s12662-025-01084-3**

[28] Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, *2022*, 1-10. DOI: **https://doi.org/10.5194/gmd-15-5481-2022**

[29] Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, *237*, 121549. DOI: **https://doi.org/10.1016/j.eswa.2023.121549**

[30] DeVries, Z., Locke, E., Hoda, M., Moravek, D., Phan, K., Stratton, A., ... & Phan, P. (2021). Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability. *The spine journal*, *21*(7), 1135-1142. DOI: **https://doi.org/10.1016/j.spinee.2021.02.007**

[31] Pan, S., Liu, Z., Han, Y., Zhang, D., Zhao, X., Li, J., & Wang, K. (2024). Using the Pearson's correlation coefficient as the sole metric to measure the accuracy of quantitative trait prediction: is it sufficient?. *Frontiers in Plant Science*, *15*, 1480463. DOI: **https://doi.org/10.3389/fpls.2024.1480463**

[32] Ma, X., Yuan, J., Chen, Y. W., Tong, R., & Lin, L. (2022). Attention-based cross-layer domain alignment for unsupervised domain adaptation. *Neurocomputing*, *499*, 1-10. DOI: **https://doi.org/10.1016/j.neucom.2022.04.086**

[33] Yang, C., Dong, Y., Du, B., & Zhang, L. (2022). Attention-based dynamic alignment and

dynamic distribution adaptation for remote sensing cross-domain scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-13. DOI: **https://doi.org/10.1109/TGRS.2022.3225589**

[34] Aljarbouh, A., Yarygina, I., Mohamed, A. P., Bystrova, N., & Tsarev, R. (2024, December). Evaluating the Effectiveness of an Online Trainer: A Paired T-Test Analysis. In *International Workshop Hybrid methods of modeling and optimization in complex systems* (pp. 338-347). Cham: Springer Nature Switzerland. DOI: **https://doi.org/10.1007/978-3-031-95649-2_29**

[35] Uddin, S., & Lu, H. (2024). Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *Plos one*, *19*(4), e0301541. DOI: **https://doi.org/10.1093/ecco-jcc/jjaf231.1458**