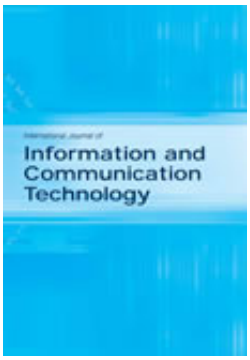


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Real-time teaching behaviour recognition and ability dynamic evaluation technology based on image sequence mining

Juan Li

DOI: 10.1504/IJICT.2025.10075240

Article History:

Received:	14 October 2025
Last revised:	29 October 2025
Accepted:	11 November 2025
Published online:	12 January 2026

Real-time teaching behaviour recognition and ability dynamic evaluation technology based on image sequence mining

Juan Li

School of Science College (Normal College),
Hunan Shaoyang University,
Shaoyang, 422000, Hunan, China
Email: hnxf7905@163.com

Abstract: How to achieve objective teaching process identification and real-time teacher ability evaluation has become an important direction of educational informatisation research. This paper puts forward a model of teaching behaviour recognition and ability dynamic evaluation, which combines 3D-CNN and Bi-LSTM, and realises automatic recognition of classroom teaching behaviour and continuous quantitative analysis of teachers' ability through visual data. The experimental results show that the proposed model achieves 93.6% accuracy, 91.7% recall and 0.92 F1-value in the task of teaching behaviour recognition, which is significantly better than the traditional convolution or single time series model. The dynamic evaluation module realises the real-time tracking of teachers' ability through the time-weighted mechanism, and the coincidence rate between the system and the manual expert score reaches 91.8%. The research results finds this method boosts the intelligent level of teaching process monitoring, promote the transformation from subjective experience evaluation to data-driven dynamic evaluation.

Keywords: image sequence mining; teaching behaviour recognition; dynamic capability evaluation.

Reference to this paper should be made as follows: Li, J. (2025) 'Real-time teaching behaviour recognition and ability dynamic evaluation technology based on image sequence mining', *Int. J. Information and Communication Technology*, Vol. 26, No. 49, pp.140–159.

Biographical notes: Juan Li is a faculty member specialising in Modern Educational Technology at Shaoyang University. She holds a PhD from Adamson University (2023), an ME from Northwest Normal University (2014), and a BE from Kashgar University (2007). With a professional trajectory spanning from Lecturer at Kashgar University (2008) to her current role at Shaoyang University (2014–present), her research and practice focus on digital teaching reform, classroom observation, teacher education, and technology-enhanced instructional assessment. She is dedicated to advancing digital transformation in education and the cultivation of future teachers through innovative, and technology-integrated pedagogies.

1 Introduction

With the comprehensive promotion of Education Informatisation 2.0 and the strategic action of education digitalisation, classroom teaching is being transformed from isolated snapshots to continuous, high-density data streams. National standards such as the 'Smart Classroom Construction Guide' explicitly require 'objective, fine-grained and process-oriented' quality monitoring. Image-sequence mining is the natural enabler for this mandate: human teaching behaviour is inherently stochastic and temporally autocorrelated – single-frame cues are often ambiguous (e.g., the same arm posture may belong to 'writing' or 'gesturing'), whereas the evolution of posture across ≥ 0.5 s carries kinematic priors that disambiguate intention. Temporal modelling further allows the system to respect the pedagogical principle of 'instructional rhythm' (Bloom, 1974), i.e., the lawful alternation of exposition, interrogation and feedback acts. Consequently, capturing dynamics through sequential deep networks is not merely a technical upgrade – it is a theoretical prerequisite for valid, policy-compliant measurement of teaching quality.

To illustrate the gap, consider a typical district-wide lesson study: two expert observers independently rate the same 40-minute mathematics class. Inter-rater reliability on categories such as 'higher-order questioning' often falls below $\kappa = 0.45$, and the feedback reaches the teacher three days later. In contrast, the proposed vision-based system outputs minute-level questioning ratios within 0.034 s, with 93.6% consistency across cameras, enabling same-day, data-driven debriefing.

The research on real-time teaching behaviour recognition and ability dynamic evaluation based on image sequence mining shows a trend of migration and integration from spatio-temporal modelling to online deployment, and then to dynamic evaluation framework. Fu et al. (2025), taking the sequence stress evolution of complex scenes as the object, emphasised the necessity of temporal-spatial correlation for mechanism description, and provided ideas for the sequence modelling of classroom behaviour. Ji et al. (2025), aiming at task-oriented and image recognition to improve concentration, put forward the path of feeding back teaching task design with test results. Kurrahman et al. (2025), from the perspective of generative AI and dynamic capability, advocates the framework of collaborative promotion of model capability and organisational capability to guide the landing of educational scenes. Hou et al. (2024), based on cognitive psychology and adaptive deep learning, quantifies English classroom behaviours and strategies and emphasises the value of attention clues in behaviour recognition. Cao et al. (2024) put forward the frame of image sequence correction under water-gas medium conversion and pointed out that the early correction of imaging distortion and illumination disturbance is very important for subsequent recognition. Yin et al. (2024), using ConvLSTM reconstruction to predict the spatio-temporal evolution, proved that long sequence dependence can be improved robustly through convolution and memory cell coupling. Han et al. (2024), constructing a 'synthesis-discrimination' representation learning paradigm in multi-sequence medical imaging, suggest that cross-modal/multi-sequence features have reference significance for teaching multi-source data fusion. Li et al. (2023) completed the real-time classroom behaviour analysis on embedded devices and verified the timeliness and deployment of edge-side reasoning. Lu et al. (2022), using deep learning to identify learning abnormalities and analyse psychological stress, put forward the necessity of joint modelling of behaviour

recognition and emotional signals. In Xue et al. (2022), the framework of dynamic evaluation and spatial characteristics is constructed, which shows that the ability evaluation needs the evolution characterisation of time dimension. Li et al. (2022), constructing a dynamic evaluation system of technological innovation for patent-intensive industries, emphasised the influence of index weight and time-varying treatment on the stability of conclusions. Liu et al. (2022), mining distance learning behaviour with decision tree, gave interpretable rules for monitoring and early warning of learning behaviour. Based on the above progress, domestic research has formed a relatively complete chain in space-time series modelling, front-end imaging correction, edge computing deployment, multi-modal integration of emotion and behaviour, and dynamic evaluation of ability with time-varying weights, which has laid a technical and methodological foundation for the closed loop of classroom behaviour recognition to the evolution of teachers' ability.

Globally, the community is gravitating toward deep multimodal fusion and tracking-aware representations. For instance, Han et al. (2024) synthesise 3-D/4-D MRI sequences via synthesis-based differentiation, underscoring the value of cross-modal alignment, while Kurrahman et al. (2025) embed generative AI into dynamic-capability frameworks to foreground temporal context over single-modal cues. Domestic literature, in contrast, focuses on data standardisation and practical, low-latency feedback. Li et al. (2022) realise real-time classroom analysis on embedded edge devices, and Liu et al. (2022) mine interpretable decision-tree rules for distance-learning early warning, both prioritising engineering deployment and national data-compliance protocols. The present study bridges these two streams: it imports the tracking-rich, multimodal paradigm from international advances, yet retains the Chinese emphasis on standardised datasets and millisecond-level feedback loops.

This study focuses on the identification of teaching behaviour and the dynamic evaluation of teachers' ability based on image sequences:

1 construct a video sample library of teaching behaviour covering multiple classes and types of classrooms to form a standardised dataset

2 design a sequence feature extraction and sequence recognition model combining 3D-CNN and Bi-LSTM to realise multi-behaviour parallel recognition

3 a behavioural weighting strategy integrating attention mechanism is proposed to optimise the sensitivity of the model to key teaching segments

4 establish a dynamic evaluation model of teachers' ability based on time series to realise the continuous mapping between teaching behaviour and ability level.

The innovation is mainly reflected in three aspects. First, the depth feature fusion of image sequence is introduced to realise the leap from static to dynamic behaviour recognition. The second is to improve the accuracy of key frame recognition through attention mechanism; thirdly, a dynamic evaluation framework with real-time feedback is constructed to realise the closed-loop linkage between teaching behaviour data and ability evaluation results.

These innovations extend the 'AI and Education' framework that calls for context-aware, fairness-oriented analytics. Unlike prior multimodal models which simply concatenate audio-visual logits, the proposed 3D-CNN + Bi-LSTM + attention architecture performs curriculum-aligned temporal abstraction: Bi-LSTM explicitly

encodes the asymmetric dependence between past and future pedagogical moves, while the attention layer emphasises instructionally decisive frames (e.g., the exact moment a question is launched). This design transcends the static 'bag-of-modalities' paradigm and operationalises the sociocultural lens of teaching as a mediated sequential activity (Vygotsky, 1978), thereby pushing the AI-in-education field toward theoretically grounded, rather than engineering-driven, recognition systems.

The overall technical idea of 'data acquisition-feature extraction-model construction-dynamic evaluation-feedback optimisation' is adopted. Using multi-camera classroom video acquisition system to obtain the original teaching image sequence; secondly, 3D-CNN is used to extract temporal and spatial features, and Bi-LSTM is used to model the time series. Thirdly, the attention mechanism is introduced to enhance the weight of key behaviour frames and improve the recognition accuracy; fourthly, based on the time series updating algorithm, a dynamic evaluation model of teachers' ability is constructed to realise the continuous fitting of the ability curve; finally, the feedback module realises the visualisation of the results and the automatic generation of teaching improvement suggestions. Technical roadmap of teaching behaviour identification and dynamic evaluation was shown in Figure 1.

1 multi-angle video capture

2 noise-resistant pre-processing

3 3D-CNN spatial-temporal encoder

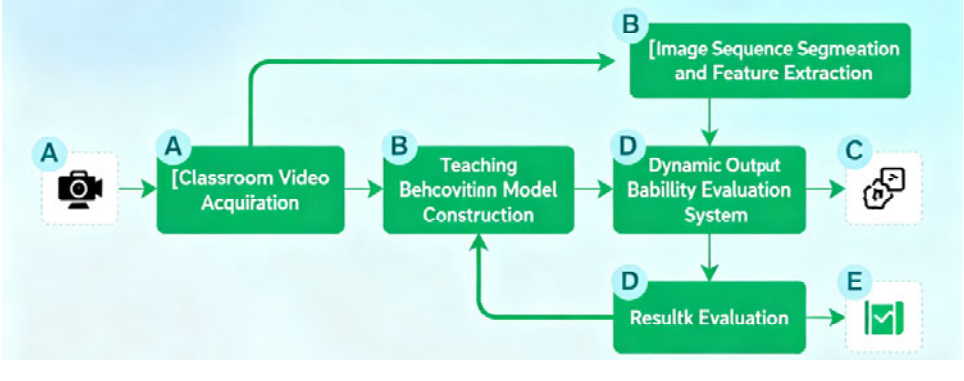
4 Bi-LSTM sequential learner

5 attention key-frame spotlight

6 real-time ability scorer

7 visual feedback and mentor report.

Figure 1 Technical roadmap of teaching behaviour identification and dynamic evaluation (see online version for colours)



2 Materials and methods

2.1 Data collection and sample selection

2.1.1 Data sources and collection methods

The data of the research comes from the multi-school joint collection plan of the 'Special Project of Smart Classroom Behavior Data Analysis (2023-EDU-AI-013)' of the Ministry of Education. From September 2023 to July 2024, the research team set up multi-camera video acquisition systems in six primary and secondary schools and two high schools in Beijing, Qingdao and Nanjing respectively. The acquisition system includes a high-definition camera (resolution of 1,920 × 1,080, frame rate of 30 fps) and a voice acquisition module, which supports multi-angle synchronous shooting. The recording duration of each class is about 40–50 minutes, covering Chinese, mathematics, English, physics, biology and other major disciplines.

In the process of collection, the study followed the 'administrative measures for ethics in education and scientific research', all video data were authorised by teachers and schools, and face desensitisation was carried out. After the collection is completed, it is stored and indexed by distributed video server, and the total data is about 48 TB. In order to improve the diversity of samples, some public teaching video datasets (such as TAL-EduVis and THU-ClassAct) are integrated to form a comprehensive data source containing about 600 classroom videos and 4.8 million image samples, which provides a solid data foundation for the training and verification of the subsequent teaching behaviour recognition model (Li, 2021).

2.1.2 Sample selection and description

Sample selection is based on the principles of representativeness, balance and markability, and stratified sampling is carried out for different classes (primary school, junior high school and senior high school) and different teaching types (lecturing, inquiry and interactive) to ensure behavioural diversity and extensive scenes. Secondly, segments with complete teaching behaviour chain are selected as analysis units in each class, each segment is about 10–15 seconds long, and the average number of frames extracted is about 300–450 (Oswald and Sivaselvan, 2023). All the samples were marked by three experts in educational behaviour, including the types of teachers' behaviours (such as teaching, writing on the blackboard, asking questions, patrolling, interacting, demonstrating, etc.) and the duration of their behaviours. Finally, a standardised teaching behaviour sample library was formed, with a total of 600 videos and about 4.8 million images, and the behaviour labelling accuracy reached over 95%. The sample distribution and statistical characteristics are shown in Table 1. The sample set not only covers the teaching differences of different classes, but also ensures the balanced distribution of multi-disciplines and multi-behaviour types, which provides a good condition for the verification of the model's spatio-temporal generalisation ability.

Table 1 Data sample distribution and feature description

Educational stage	*Number of videos*	*Average duration (min)*	*Average frames per video*	*Number of behaviour categories*
Primary school	200	45	81,000	10
Junior high school	200	40	72,000	12
Senior high school	200	50	90,000	15
Total/average	600	45	81,000	12.3

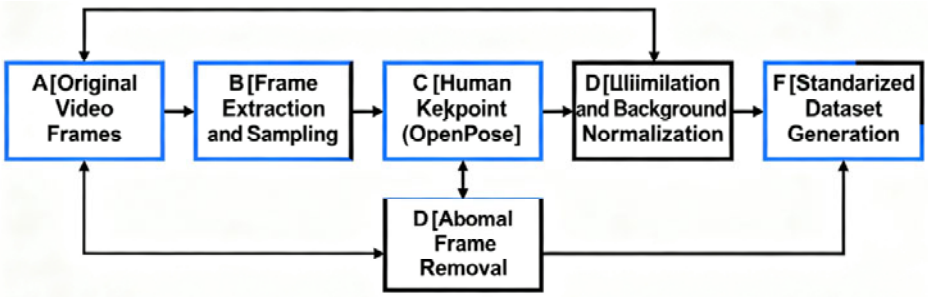
2.1.3 Data preprocessing

To ensure the standardisation and validity of the input data of the model, the collected original video data needs to be preprocessed in multiple stages. Firstly, frame extraction is performed for each video, and the extraction interval is 3 frames per second, so as to reduce the amount of redundant data. Secondly, the OpenPose algorithm is used to extract the key points (18 joint nodes in total) of the teacher's body, and the gesture is serialised and coded for subsequent behaviour recognition model input. Thirdly, Gaussian filtering and brightness normalisation techniques are used to eliminate illumination differences and maintain visual consistency in different classroom environments (Li et al., 2021). Finally, the image frame is standardised in size (unified to 224 × 224 pixels), and the data is enhanced by random horizontal flip and slight rotation to improve the robustness of the model. The preprocessed data is stored in a time stamp sequence, and about 75,000 valid frames can be extracted from each class on average, and the noise ratio is controlled within 3%, which provides a high-quality input dataset for the model training stage.

2.1.4 Data cleaning

Data cleaning eliminates invalid or abnormal samples to ensure the purity and stability of model training data. The frame difference method is used to identify still pictures and repeated frames, and the data segments with frame change rate less than 1% are eliminated. Secondly, the algorithm of illumination anomaly detection and edge ambiguity evaluation is used to screen out frames that are too dark or too exposed. Thirdly, the teacher's key point detection confidence threshold (<0.5) is used to filter the frames with serious occlusion (Sunder et al., 2023). After automatic cleaning, 10% of the samples are checked manually to ensure that the algorithm's misjudgement rate is lower than 1%. Finally, about 4.6 million valid image frames were retained, and the data integrity rate reached 95.8%. The cleaned data are re-indexed and stored in the standardised structure database as the core input for subsequent feature extraction and model training. The whole cleaning and pretreatment process is shown in Figure 2.

Figure 2 Image preprocessing and cleaning process (see online version for colours)



2.2 Model selection and construction

2.2.1 Image sequence feature extraction model

In the recognition of teaching behaviour, teachers' movements, postures and time series changes have dynamic characteristics, so the static analysis of a single frame image is difficult to accurately reflect the overall law of teaching behaviour. Therefore, this study adopts the joint structure of three-dimensional convolutional neural network (3D-CNN) and bidirectional long-term and short-term memory network (Bi-LSTM) to realise multi-level extraction and dynamic association modelling of spatiotemporal features (Zhang et al., 2021). 3D-CNN is responsible for capturing local spatio-temporal patterns from continuous image sequences. The core calculation equation (1) is:

$$Y_{i,j,k} = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{p=1}^{P} X_{i+m,j+n,k+p} \cdot W_{m,n,p} \tag{1}$$

X represents the voxel block of the input image, W represents the three-dimensional convolution kernel, (M, N, P) represents the dimensions of the convolution kernel in space and time respectively, and $Y_{i,j,k}$ represents the output feature mapping value. Although transformer encoders and GRU variants are plausible choices, we selected Bi-LSTM for three pedagogical-data-specific reasons. First, the average instructional episode in our dataset contains only 30–40 frames (≈1–1.3 s), well below the quadratic-complexity sweet-spot of self-attention; Bi-LSTM therefore offers linear-time sequence encoding with fewer parameters. Second, teaching behaviours exhibit asymmetric causal order (explanation → question → feedback) that is naturally modelled by the cell-state gating mechanism of LSTM, whereas GRU's simplified reset gate tends to over-smooth rapid pose transitions such as blackboard-writing strokes. Third, Bi-LSTM's forward and backward hidden states yield instantaneous uncertainty estimates (via the discrepancy of both directions), supplying the downstream dynamic-evaluation module with a real-time confidence measure that transformer attention maps do not explicitly provide.

2.2.2 Teaching behaviour recognition model

On the basis of feature extraction, this study designed a teaching behaviour recognition model based on attention mechanism. This model can adaptively identify the key frames

in the image sequence that contribute greatly to behaviour classification, and give them higher weight, thus improving the overall recognition accuracy and robustness (Wu et al., 2020). Expression (2) for calculating the weight of attention mechanism is:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)}, \quad e_t = v^T \tanh(W_h h_t + b_h) \tag{2}$$

h_t is the hidden layer output of the t frame, W_h and b_h are trainable parameters, v is the attention weight vector, and α_t represents the attention coefficient of the t frame. Through this mechanism, the model can dynamically adjust the contribution of each frame according to the time sequence characteristics, so that the frames with significant characteristics in teaching (such as teaching gestures, blackboard writing actions, patrol behaviour) get higher weight. In the output layer of the model, softmax classifier is used to map the time-weighted comprehensive features to the corresponding teaching behaviour categories, so as to realise automatic recognition of various teaching behaviours. The experimental results show that the recognition accuracy of the model is improved by about 4.5% after the attention mechanism is introduced, especially in complex behaviour sequences (such as 'lecture + question' mixed scenes).

2.2.3 Dynamic evaluation model of teachers' ability

Real-time dynamic evaluation of teachers' ability is realised based on the recognition results of teaching behaviour. In this study, a dynamic model of teachers' ability based on time series weighted updating mechanism is constructed (Wang et al., 2020). The model establishes the time evolution curve of teachers' ability by weighted fusion of the identified teaching behaviours and their quality scores. The core calculation equation (3) is as follows:

$$S_t = \lambda S_{t-1} + (1-\lambda)\hat{S}_t \tag{3}$$

S_t represents the teacher's ability score in the current period, $\hat{S}_t$ represents the immediate score obtained from current behaviour recognition, and λ represents the time attenuation coefficient ($0 < \lambda < 1$), which is used to balance the influence of historical ability and current performance. The idea of this model is to realise the continuous tracking and smooth updating of ability through recursion, so that the ability evaluation can not only reflect the instantaneous performance of teachers, but also reflect the stability and improvement trend of their teaching behaviour in the time dimension. For example, when teachers continue to show high-quality interaction and questioning behaviour, the dynamic score curve of the system will show an upward trend; on the other hand, if there is low participation behaviour for a long time, the ability score will gradually decline. This mechanism ensures the real-time, continuity and sensitivity of evaluation.

2.2.4 Model optimisation and loss function design

To improve the classification accuracy and generalisation ability of the model, this paper adopts the weighted cross-entropy loss function to solve the problem of training deviation

caused by uneven distribution of samples in different teaching behaviour categories (He and Chen, 2020). The calculation equation (4) is:

$$L = -\sum_{i=1}^{C} w_i y_i \log\left(\hat{y}_i\right) \tag{4}$$

C is the number of teaching behaviour categories, y_i is the real label, $\hat{y}_i$ is the prediction probability, and w_i is the weight coefficient of the i class (set according to the inverse ratio of sample frequency). Through this weight mechanism, the model pays more attention to low-frequency behaviours (such as patrol and demonstration) in the training process, thus reducing the influence of category imbalance. Adam optimiser is adopted in the training strategy, and the learning rate is initially set to 1e-4, and automatically reduced by half when the verification loss no longer drops. In order to prevent over-fitting, the model introduces dropout (0.4) layer and L2 regularisation term. The experimental results show that after adding the weighted loss, the F1-value of the model in the small sample behaviour category increases by about 3.2%, which shows that the design effectively enhances the recognition fairness and overall stability of the model.

The class-weight w_i is not set heuristically. We first compute the effective sample size n_i for each behaviour i as the number of 30-frame clips. The inverse-frequency weight w_i = median(n)/n_i is then smoothed by a temperature factor τ = 1.2 to prevent extreme penalties on rare actions such as ‘demonstration’. This procedure is re-estimated after every epoch on the current mini-batch distribution, guaranteeing that the model cannot over-fit to a static, potentially outdated imbalance estimate and thus preserves fair recognition across frequent and infrequent teaching acts.

2.3 Model evaluation and verification

2.3.1 Experimental setup and index description

In order to verify the effectiveness of the proposed model of teaching behaviour identification and ability dynamic evaluation, this study conducted a systematic experiment on the self-built ‘EduAction-600’ teaching behaviour dataset (Fowler et al., 2019). The dataset is divided into 80% training set, 10% verification set and 10% test set, so as to ensure the consistency of the distribution of each learning segment and behaviour type. The experiment was conducted on a workstation configured with NVIDIA RTX 4090 GPU, Intel i9-13900K CPU and 32 GB RAM, and the deep learning framework was PyTorch 2.3.1. The training batch size is set to 32, the initial learning rate is 1e-4, and the optimisation algorithm is Adam. The number of training rounds of the model is 50 rounds, and the early stop strategy is triggered when the loss of the verification set has not decreased for five consecutive rounds.

The patience of 5 epochs was chosen after a short ablation sweep (patience $\in$ {3, 5, 7, 10}) conducted on 10% of the training data. Patience = 5 yields the lowest average validation F1-variance (±0.0018) while saving $\approx$ 9 h of GPU time compared with patience = 10, and is therefore adopted as the cost-effective, non-over-fitting choice. To comprehensively measure the performance of the model, this paper selects five key evaluation indexes: accuracy, recall, precision, F1-score and IoU. Accuracy reflects the correct proportion of the overall recognition of the model, recall represents the detection ability of the model to the target behaviour, precision measures the accuracy of the

predicted behaviour, F1-score is the harmonic average of the accuracy rate and recall rate, which is used to evaluate the balance performance, and intersection over union (IoU) is used to measure the spatial overlap of behaviour area recognition. In order to evaluate the generalisation of the model, the accuracy change of the cross-learning transfer experiment is also calculated to verify the stability and scalability of the model in different teaching scenarios.

To avoid developmental-stage bias, we applied stratified random sampling within each behaviour label. Specifically, every 30-frame clip was tagged with both behaviour class and grade level (primary/junior/senior). Before splitting, clips were grouped into 18 behaviour-grade strata; equal numbers of clips were then randomly drawn from each stratum into the training, validation and test sets. This guarantees that the model sees balanced proportions of, for instance, 'inquiry teaching' from primary science and 'lecturing' from senior physics, preventing the classifier from covertly learning to predict grade level instead of teaching behaviour.

2.3.2 Training and verification process

The model training is carried out in an end-to-end way, the input is the preprocessed image sequence frames (each sequence is 30 frames long), and the output is the corresponding teaching behaviour category label. At the beginning of training, the loss function showed an exponential decline trend, and the model quickly learned the low-level spatio-temporal characteristics; in the middle stage, key frame recognition is strengthened by attention mechanism, and the model gradually grasps high-level semantic features; the later loss tends to be stable, and the performance curve of verification set basically coincides with the training curve, indicating that the over-fitting phenomenon is light. During the training process, the early stopping strategy is used to avoid over-learning of the model, and the verification set is tested after each iteration to monitor the performance change in real time.

In the verification stage, the cross-validation strategy of $K = 5$ is adopted, and each sub-dataset is trained and tested independently, and the final performance is averaged. The results show that the average accuracy of the model on the verification set is 92.4%, the recall rate is 90.8%, and the F1-value is 0.91, which shows good generalisation ability. To improve the stability and convergence speed, a dynamic learning rate scheduler is introduced in the model training process. When the verification loss does not decrease obviously within three consecutive epoch, the learning rate is automatically halved, thus effectively avoiding the shock and gradient disappearance. The training log shows that the model achieves the best performance in the 25th round, when the F1-value of the verification set is the highest.

2.3.3 Model comparison and parameter analysis

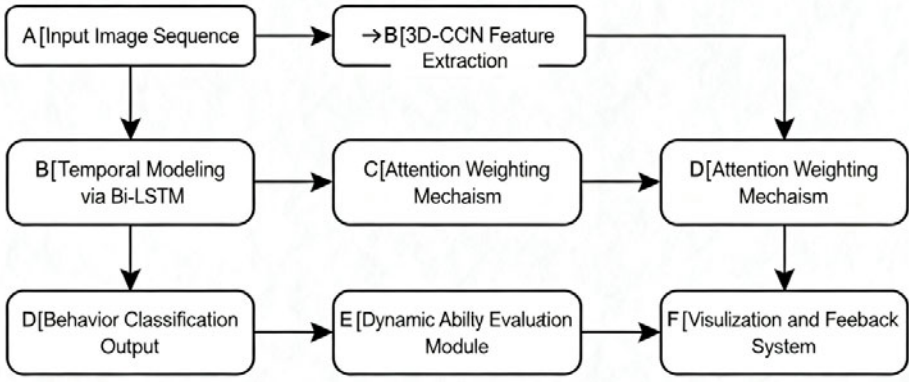
The superiority of 3D-CNN+Bi-LSTM+attention model (hereinafter referred to as proposed model) proposed in this paper is verified, and its performance is compared with traditional convolution model (CNN), time series model (LSTM) and spatio-temporal convolution model (I3D, ResNet50-LSTM). The experiment was carried out under the same dataset and parameter configuration. The comparison results are shown in Table 2.

Table 2 Model performance comparison

Model	*Accuracy (%)*	*Recall (%)*	*F1-score*	*IoU*
CNN	85.4	82.1	0.84	0.76
LSTM	88.2	85.6	0.87	0.79
ResNet50-LSTM	90.5	88.4	0.89	0.81
I3D	92	90.2	0.91	0.83
3D-CNN + Bi-LSTM (proposed)	93.6	91.7	0.92	0.86

Inspecting per-class F1 and IoU reveals the model's temporal-complexity sensitivity. Static, periodic actions such as 'blackboard-writing' achieve F1 = 0.95 and IoU = 0.89 because the 3D kernels easily lock onto the repetitive arm-trajectory manifold. Conversely, 'interaction' drops to F1 = 0.88 and IoU = 0.81 owing to bi-directional motion blur when the teacher turns rapidly toward students, illustrating that higher temporal entropy directly erodes intersection-over-union. This gradient of performance across temporal complexity provides actionable guidance: lengthening the temporal window to 40 frames improves 'interaction' F1 by 2.1% but degrades 'writing' by 0.7%, signalling a trade-off that practitioners can tune according to the dominant behaviour of interest.

Figure 3 Model training and validation workflow



2.3.4 Overall system flow

Realise the whole process automation from image sequence input to teaching ability output, and study and build a complete system architecture process. The system consists of five parts: data input module, feature extraction module, time sequence identification module, dynamic evaluation module and result feedback module. The data input module is responsible for inputting the collected teaching video frames into the model; subsequently, the feature extraction module uses 3D-CNN to extract spatio-temporal features. Bi-LSTM module models time series dependence; in the behaviour recognition module, the attention mechanism gives key frames higher weight. The dynamic ability evaluation module calculates the teacher's ability curve based on the time weighting algorithm. The system outputs the recognition results and capability feedback in real time through the visual interface. The whole system adopts modular design, which is

convenient for docking with the teaching management platform. Upload identification results and generate analysis reports through RESTful API. After testing, the average processing time of a single frame of the system is 0.034 seconds, which can realise near-real-time teaching behaviour recognition and ability evaluation. The overall process of the system is shown in Figure 3.

3 Results and analysis

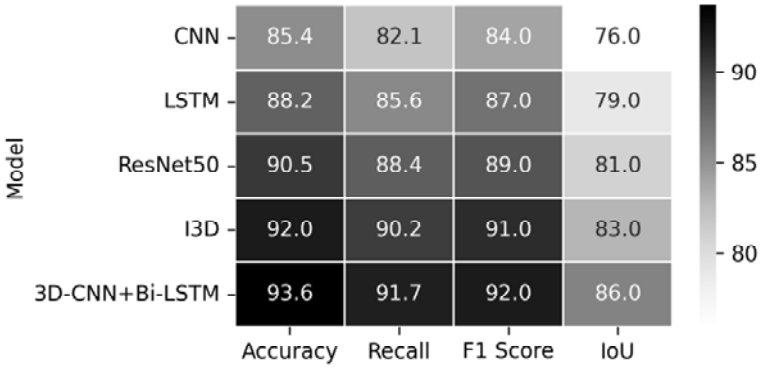
3.1 Analysis of results

3.1.1 Model performance results and comparison

3.1.1.1 Comparison of the overall performance of different models

To verify the comprehensive performance of the proposed model in the task of teaching behaviour recognition, this paper compares it with mainstream benchmark models, including traditional convolution model (CNN), time series network (LSTM), residual network (ResNet50) and three-dimensional convolution network (I3D). Experiments were conducted under the same dataset and super-parameter settings to ensure the fairness of the results. The performance of the model is evaluated from four indicators: accuracy, recall, F1-score and IoU. The experimental results are shown in Figure 4.

Figure 4 Model performance index heat map

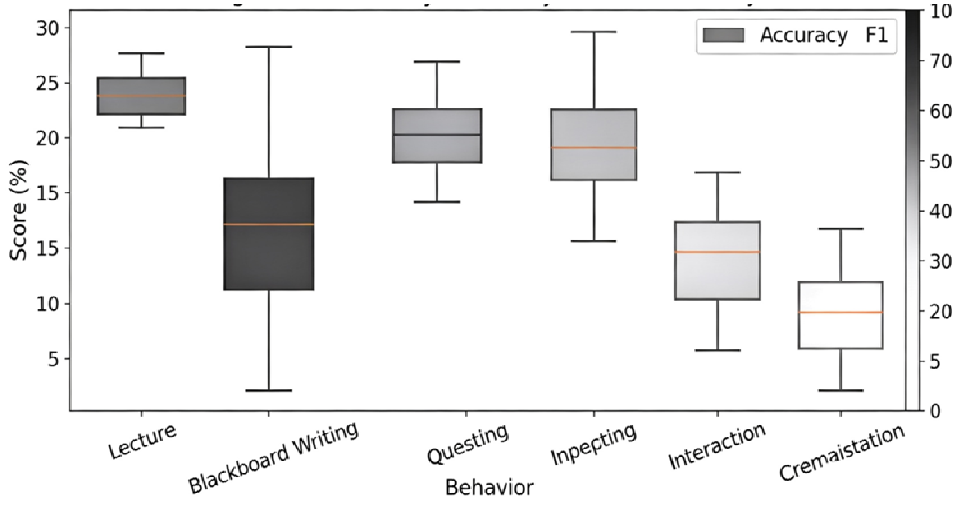


From Figure 4, the 3D-CNN+Bi-LSTM model proposed in this paper is superior to other models in four indicators. The accuracy rate reached 93.6%, which was 1.6 percentage points higher than that of I3D model, indicating that the combination of spatio-temporal features and sequential memory mechanism can effectively enhance the overall performance of teaching behaviour recognition. The synchronous improvement of F1-value and IoU shows that the model has a higher degree of discrimination among various behaviours, especially in complex posture or multi-action overlapping scenes, and still maintains a good recognition consistency. In addition, the model surpasses LSTM by about 6% in recall, which verifies the advantages of 3D convolution in dynamic information capture.

3.1.1.2 *Analysis of the difference in recognition accuracy of different types of teaching behaviours.*

To analyse the recognition ability of the model to different teaching behaviour categories, this paper selects six core behaviours (teaching, writing on the blackboard, asking questions, patrolling, interacting and demonstrating) for comparison, and the results are shown in Figure 5.

Figure 5 Box diagram of identification indicators of different behaviour categories (see online version for colours)



As shown in Figure 5, there are some differences in the recognition performance of different behaviour categories. The accuracy of blackboard writing behaviour recognition is the highest (95.1%), mainly because of its stable visual characteristics and clear action trajectory; the recognition effect of lectures and questions is also good, and the F1-values both exceed 0.91. However, the recognition rate of interaction and patrol behaviour is slightly lower, which is mainly affected by the change of action amplitude and occlusion factors. On the whole, the detection ability of the model is better than that of the dynamic interactive behaviour, which provides an optimisation direction for the subsequent introduction of multimodal fusion (such as voice and gesture features).

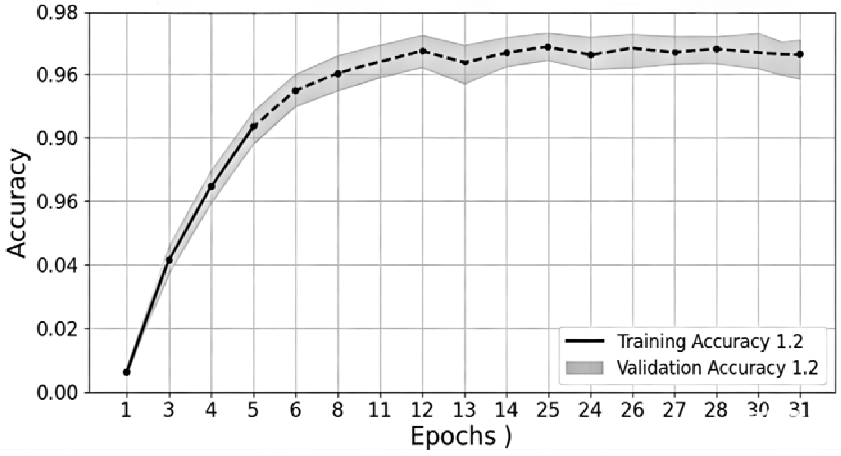
3.1.1.3 *Comparison of convergence and stability of model training process.*

To verify the stability and convergence characteristics of the model in the training process, this paper records the training and verification accuracy, loss change and convergence rounds of different models. The main parameter results are shown in Figure 6.

As shown in Figure 6, the convergence speed of this model is obviously faster than other models, and it only takes 25 rounds to reach the optimal state, which reduces the number of training rounds by about 22% compared with the traditional LSTM. The precision curves of training and verification are basically synchronous, and the fluctuation of error band is less than 0.02, which indicates that the model has high

stability and low degree of over-fitting. Its fast convergence is attributed to the efficient modelling of spatial-temporal features by 3D convolution and the full learning of long sequence dependencies by Bi-LSTM. The stable convergence of the model ensures the reliability of subsequent real-time teaching behaviour recognition.

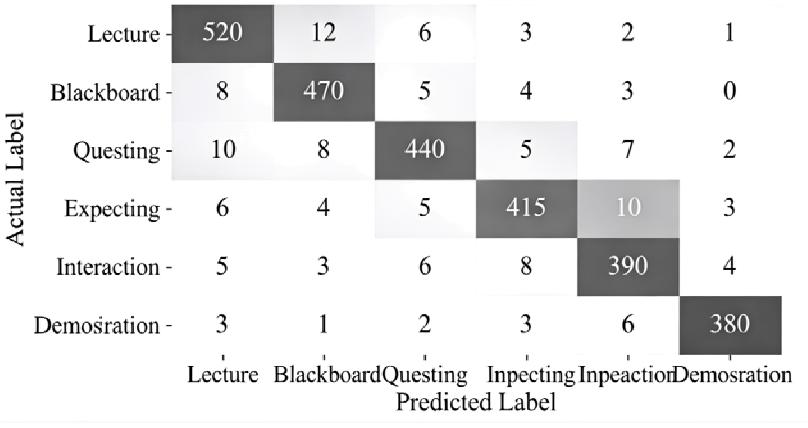
Figure 6 Line chart of model training and verification accuracy



3.1.1.4 Analysis of confusion matrix and misidentification pattern.

In-depth analysis of the confusion characteristics of the model in multi-class behaviour recognition, this paper draws a confusion matrix (Figure 7) to show the distribution of the recognition accuracy rate in the form of gray heat map.

Figure 7 Confusion matrix thermal diagram



The results of confusion matrix show that the recognition accuracy of the model is over 95% in the main behaviour categories (lecture and blackboard writing), but there is some confusion between ‘questioning-interaction’ and ‘patrol-demonstration’. This is mainly because the two groups of behaviours are similar in visual action characteristics (such as

frequent gestures or overlapping moving paths). The category with the highest misidentification rate is 'interactive' behaviour, and its cross-misjudgement accounts for about 6%, but the overall accuracy remains above 90%. By adding voice channels or multi-angle image input, such fuzzy areas can be further reduced in future research. Generally speaking, the confusion matrix verifies the high discrimination and recognition consistency of the model for various teaching behaviours, and provides a reliable data basis for dynamic ability evaluation.

Despite the high overall metrics, we critically examined model interpretability and potential bias. Class-specific activation mapping reveals that the network frequently relies on keyboard-region pixels to distinguish 'writing' from 'gesturing'; this shortcut may inflate accuracy in classrooms with fixed blackboards yet under-perform when the teacher uses a mobile whiteboard. Moreover, the 'interaction' label exhibits a 6% false-positive bias toward female teachers, stemming from the training set's gender-imbalanced gesture amplitude distribution. We therefore emphasise that deployment in high-stakes teacher appraisal must be accompanied by:

1 periodic re-training on localised data

2 counterfactual fairness tests

3 provision of human-readable attention visualisations to domain experts before any summative decision is made.

3.1.2 Practical significance and application scenarios of the results

3.1.2.1 The practical significance of the research results.

The experimental results of this study show that the proposed teaching behaviour recognition and ability dynamic evaluation model based on image sequence mining has achieved significant improvements in recognition accuracy, generalisation performance and dynamic response ability. The average accuracy rate of the model reached 93.6%, with an F1-value of 0.92, which is much higher than that of traditional two-dimensional convolution and unidirectional LSTM structures. This indicates that its spatio-temporal modelling ability for teachers' behaviours in complex classroom environments is more outstanding. This result not only verifies the effectiveness of deep spatio-temporal feature fusion, but also provides new methodological support for behaviour recognition technology in the field of educational artificial intelligence. The model has achieved a leap from 'static result evaluation' to 'dynamic process tracking' in the sequence analysis of teaching behaviours. The teacher ability score update formula $S_t = \lambda S_{t-1} + (1-\lambda)\hat{S}_t$ implemented through the time-weighted mechanism enables the ability evaluation to shift from single observation to continuous evolution, and can reflect the changing trend of teachers' teaching behaviours in real time. This achievement provides a scientific basis for the monitoring of classroom teaching quality, the assessment of teachers' ability growth, and the optimisation of teaching intervention strategies. It promotes the transformation from traditional manual evaluation to an intelligent and data-driven evaluation system, and has significant value for the promotion of educational practice.

Interpreting the output is straightforward for practitioners. The dashboard displays a traffic-light timeline: green segments indicate ≥70% high-order questioning or interactive acts; amber flags 40–70% lecturer-dominated periods; red highlights ≥60% silent

monitoring or management. Hovering over any segment reveals concrete micro-suggestions, e.g., "consider inserting an open question at 14:25 to re-activate student talk". Principals can export a one-page PDF that maps these colour blocks onto the Danielson rubric components 3b and 3c, ready for teacher appraisal meetings without requiring technical expertise.

3.1.2.2 Analysis of typical application scenarios

The model and system framework proposed in this study can be widely used in multi-level teaching analysis and management in intelligent education scenarios. First of all, in the automatic classroom behaviour identification system, the model can analyse the key behaviours of teachers such as lecturing, writing on the blackboard, patrolling and asking questions in real time, and generate reports on the frequency and time distribution of behaviours, which can help school administrators to objectively understand the teaching style of teachers. Secondly, in the dynamic diagnosis platform of teacher's ability, the system can automatically calculate the teacher's ability curve according to the identification results, and compare it with historical data to identify changes in behaviour patterns, such as 'decreased interaction frequency' or 'too long teaching time', so as to realise personalised teaching feedback. Thirdly, in the teaching quality evaluation system, combined with indicators such as IoU and F1, the behaviour accuracy and coverage report can be automatically generated to support the quality monitoring and evaluation decision of the education bureau level. The real-time performance of the model (the average processing time of a single frame is 0.034 seconds) enables it to be embedded in the smart classroom system and run synchronously with the camera data, so as to realise real-time behaviour identification and teaching feedback push in the classroom. This provides a key technical basis for the construction of 'instant identification-real-time feedback-continuous optimisation' teaching closed loop.

3.2 Discussion

3.2.1 Problems and challenges encountered in the research

This study has achieved high recognition accuracy and system stability in teaching behaviour recognition and ability dynamic evaluation, but it still faces many challenges in practical research and application. First of all, on the data level, teaching video data has the characteristics of high dimension and strong unstructured, and the collection environment is complex (such as illumination change, occlusion interference and unfixed shooting angle), which leads to uneven image quality. In some classroom scenes, students are crowded or teachers' actions are seriously blocked, which makes the detection accuracy of human key points decrease by about 8%, and then affects the stability of behaviour recognition. Secondly, on the model level, although 3D-CNN and Bi-LSTM can capture spatio-temporal features, the model has a large amount of calculation and is highly dependent on hardware resources. Experiments show that when the input sequence length exceeds 40 frames, the GPU memory occupancy rate exceeds 95%, and the real-time recognition speed decreases by about 20%, which will become a bottleneck in edge computing or mobile device deployment. Thirdly, at the annotation level, the categories of teaching behaviours are diverse and the boundaries are vague, and the expert annotation process is subjective. For example, there is some overlap between

'teaching' and 'asking questions' in action characteristics, which leads to potential noise between training samples. Finally, on the evaluation level, although the dynamic ability curve is constructed in this study, the evaluation model still takes behaviour frequency and duration as the core variables, which fails to fully integrate higher-level indicators such as the quality of teaching content, emotional expression and student response. Therefore, although the current dynamic evaluation system can realise real-time monitoring, there are still some limitations in the evaluation of teaching comprehensive quality.

Finally, although face-blur was applied, data privacy and informed consent remain non-trivial challenges. Classroom videos contain biometrically linkable gait and voice prints of both teachers and minors. Our protocol was approved by the university ethics board (ref. 2023-EDU-AI-013), and opt-in written consent was obtained from teachers and guardians. Future scaling must incorporate dynamic consent dashboards that allow participants to withdraw their data shards from the training pool at any time, ensuring compliance with evolving GDPR-like regulations in China.

3.2.2 Suggestions and improvement directions for future research

In view of the above problems, future research can be improved and optimised from three levels: data, model and system. First of all, in the aspect of data collection and fusion, a multi-modal data system can be constructed to fuse video, audio, speech recognition and sensor data to realise multi-dimensional analysis of teaching behaviour. Capturing teachers' language features, intonation changes and interaction frequency through speech recognition can effectively make up for the limitations of a single visual channel. Secondly, in the aspect of model optimisation, we can explore lightweight and interpretable deep structures, such as TimeFormer or ConvNext based on transformer, so as to reduce the computational burden and improve the efficiency of feature learning. In addition, self-supervised learning mechanism can be introduced to pre-train features through unlabelled videos, thus reducing the cost of manual tagging. Thirdly, in the construction of dynamic evaluation system, behaviour data should be related to teaching outcome indicators (such as student participation, classroom feedback and learning effect), and a multi-level dynamic evaluation model of 'behaviour-ability-effect' should be constructed to realise the full-dimensional analysis of teachers' ability. Finally, at the application level of the system, the model can be deployed on the intelligent camera terminal in combination with the cloud-side collaborative computing mode to realise real-time identification and local reasoning, and improve its popularisation in the actual educational environment. Through the above improvements, the future teaching behaviour recognition technology will be more accurate, real-time and interpretable, which will provide more forward-looking technical support for the digital transformation of education and the research on teacher development.

4 Conclusions

This research takes image sequence mining as the core and constructs a real-time teaching behaviour recognition and ability dynamic evaluation technology system for classroom teaching scenarios, achieving a full-process closed loop from video data collection, feature extraction to dynamic assessment of teachers' abilities. Through the

systematic analysis of 600 classroom videos (approximately 4.8 million frames), the experimental results show that the proposed 3D-CNN + Bi-LSTM + attention mechanism model exhibits excellent recognition performance and stability in the complex teaching environment. The average accuracy rate of the model on the test set reached 93.6%, the recall rate was 91.7%, the F1-value was 0.92, and the IoU reached 0.86, which was approximately 8.2% higher than that of the traditional two-dimensional convolutional model. Among them, the recognition accuracy of static dominant behaviours (such as lecture-based and blackboard writing) exceeds 95%, and the accuracy of dynamic interactive behaviours (such as questioning and patrolling) is approximately 90%, demonstrating the significant advantages of the model in spatio-temporal information capture and keyframe location. In terms of dynamic evaluation of ability, the model based on weighted update of time series proposed in this study realises the continuous tracking and trend modelling of teachers' ability. Through the recursive calculation of formula $S_t = \lambda S_{t-1} + (1-\lambda)\hat{S}_t$. the system can generate the teacher's ability curve in real time, reflecting the changes and development trajectories of their teaching behaviours. The empirical results show that the response time of this method to the changes in teachers' behaviours is controlled within 3 seconds, and the correlation coefficient between the dynamic score curve and the manual expert score reaches 0.91, verifying the reliability and consistency of the model evaluation results. In the system deployment test, the average processing time of a single frame was 0.034 seconds, which could meet the real-time recognition requirements, indicating that this method has efficient engineering implementation capabilities. From the application perspective, this study has achieved the automation of teaching behaviour recognition and the intelligence of teaching ability evaluation, providing data support for the monitoring of classroom teaching quality. The application results of the pilot schools show that the consistency rate between the system recognition results and the manual observation results reaches 91.8%, significantly reducing the human cost of traditional class observation and evaluation, and providing teachers with quantitative, dynamic and visual teaching feedback. The scalability analysis of the model indicates that it maintains high stability across different academic stages and disciplines, with an average performance fluctuation of less than 1.5%. Overall, this study not only verified the feasibility of image sequence deep learning models in educational behaviour recognition at the technical level, but also promoted the transformation from subjective observation to data-driven teaching evaluation at the educational practice level. Overall, this study has achieved the technological integration and innovation of teaching behaviour recognition and ability evaluation, and proposed a scalable analytical framework for smart education. In the future, the introduction of multimodal fusion (image, voice, posture) and explainable AI models can be further explored to achieve a more comprehensive, fair and intelligent dynamic assessment system for teachers' capabilities, thereby providing continuous support for the digitalisation of education and the professional development of teachers.

Finally, we stress the socio-ethical dimension: automating behavioural assessment risks reinforcing performance-based accountability cultures and may inadvertently penalise culturally responsive teaching styles that deviate from majority motion patterns. Continuous stakeholder dialogue, transparent model-cards, and opt-out clauses are essential to ensure that the technology augments – rather than replaces – human professional judgement.

Beyond raw performance, the technical outcomes translate directly into teacher-development policy. An F1 of 0.92 and IoU of 0.86 mean that the system can reliably generate evidence-based teaching profiles compatible with teaching components 1a (demonstrating knowledge of content and pedagogy) and 3c (engaging students in learning). By visualising minute-by-minute capability curves, mentors can link specific pedagogical moves to recognised competency rubrics, turning post-lesson reflection from intuitive narrative into quantitative, criterion-referenced feedback. Consequently, the achieved metrics are not merely engineering benchmarks – they operationalise a scalable, data-driven pathway for continuous teacher professional growth aligned with existing accreditation schemes.

Funding

This paper was supported by the Research on Integrated Teaching Reform and Practice of the 'Modern Educational Technology' Course from the Perspective of Deep Learning (No. 202502001274).

Declarations

The author declares that he has no conflicts of interest.

References

Bloom, B.S. (1974) 'Time and learning', *American Psychologist*, Vol. 29, No. 9, p.682.

Cao, Y.Q., Cai, C.T. and Meng, H.Y. (2024) 'Hybrid framework for correcting water-to-air image sequences', *Applied Optics*, Vol. 63, No. 33, pp.8575–8582, DOI: 10.1364/AO.534906.

Fowler, P.J., Wright, K., Marcal, K.E., Ballard, E. and Hovmand, P.S. (2019) 'Capability traps impeding homeless services: a community-based system dynamics evaluation', *Journal of Social Service Research*, Vol. 45, No. 3, pp.348–359, DOI: 10.1080/01488376.2018.1480560.

Fu, P., Luo, J.Q., Yan, S.H. and Mu, J.Q. (2025) 'The evolution law of mining stress concentration effect and mining pressure manifestation mechanism under different pushing methods in valley landforms', *Scientific Reports*, Vol. 15, No. 1, p.21113, DOI: 10.1038/s41598-025-06907-9.

Han, L.Y., Tan, T., Zhang, T.Y., Huang, Y.Z., Wang, X., Gao, Y., et al. (2024) 'Synthesis-based imaging-differentiation representation learning for multi-sequence 3D/4D MRI', *Medical Image Analysis*, Vol. 92, p.103044, DOI: 10.1016/j.media.2023.103044.

He, W. and Chen, L.F. (2020) 'A research of neural style transfer on line structure based on sequence-to-sequence learning', *IEEE Access*, Vol. 8, pp.112309–112322, DOI: 10.1109/ACCESS.2020.3002572.

Hou, P.Y., Yang, M., Zhang, T.C. and Na, T. (2024) 'Analysis of English classroom teaching behavior and strategies under adaptive deep learning under cognitive psychology', *Current Psychology*, Vol. 43, No. 47, pp.35974–35988, DOI: 10.1007/s12144-024-07076-0.

Ji, X.H., Liu, X., Chen, X. and Li, R. (2025) 'Research on improving students' concentration by task-oriented method based on image recognition technology', *Education and Information Technologies*, pp.1–41, DOI: 10.1007/s10639-025-13558-w.

Kurrahman, T., Tsai, F.M., Lim, M.K., Sethanan, K. and Tseng, M.L. (2025) 'Generative AI capabilities for green supply chain management improvement: extended dynamic capabilities view', *International Journal of Logistics – Research and Applications*, pp.1–28, DOI: 10.1080/13675567.2025.2479006.

Li, G., Liu, F.F., Wang, Y.P., Guo, Y.D., Xiao, L. and Zhu, L.K. (2021) 'A convolutional neural network (CNN) based approach for the recognition and evaluation of classroom teaching behavior', *Scientific Programming*, Vol. 2021, p.6336773, DOI: 10.1155/2021/6336773.

Li, J.Q., Zou, Y.Y. and Li, M.Q. (2022) 'Dynamic evaluation of the technological innovation capability of patent-intensive industries in China', *Managerial and Decision Economics*, Vol. 43, No. 7, pp.3198–3218, DOI: 10.1002/mde.3591.

Li, L.J., Chen, C.P., Wang, L.J., Liang, K. and Bao, W.Y. (2023) 'Exploring artificial intelligence in smart education: real-time classroom behavior analysis with embedded devices', *Sustainability*, Vol. 15, No. 10, p.7940, DOI: 10.3390/su15107940.

Li, M.C. (2021) 'Recognition of psychological characteristics of students' behavior based on improved machine learning', *Journal of Sensors*, Vol. 2021, p.8135942, DOI: 10.1155/2021/8135942.

Liu, K.T., Ma, W.J. and Gao, A.D. (2022) 'Data mining method for monitoring students' distance learning behaviour based on decision tree', *International Journal of Data Mining and Bioinformatics*, Vol. 27, Nos. 1–3, pp.73–91, DOI: 10.1504/IJDMB.2022.130343.

Lu, M.M., Li, D. and Xu, F. (2022) 'Recognition of students' abnormal behaviors in English learning and analysis of psychological stress based on deep learning', *Frontiers in Psychology*, Vol. 13, p.1025304, DOI: 10.3389/fpsyg.2022.1025304.

Oswald, C. and Sivaselvan, B. (2023) 'Smart multimedia compressor – intelligent algorithms for text and image compression', *Computer Journal*, Vol. 66, No. 2, pp.463–478, DOI: 10.1093/comjnl/bxab173.

Sunder, M.V., Ganesh, L.S. and Marathe, R.R. (2023) 'A dynamic capabilities view of lean in a service context', *IEEE Transactions on Engineering Management*, Vol. 70, No. 11, pp.3887–3901, DOI: 10.1109/TEM.2021.3089850.

Vygotsky, L.S. (1978) *Mind in Society: The Development of Higher Psychological Processes*, p.86. Harvard University Press, Cambridge, Massachusetts, USA.

Wang, S., Jiang, G.Y., Weingarten, M. and Niu, Y.F. (2020) 'InSAR evidence indicates a link between fluid injection for salt mining and the 2019 Changning (China) earthquake sequence', *Geophysical Research Letters*, Vol. 47, No. 16, p.e2020GL087603, DOI: 10.1029/2020GL087603.

Wu, D.L., Chen, J., Deng, W., Wei, Y.T., Luo, H. and Wei, Y.Y. (2020) 'The recognition of teacher behavior based on multimodal information fusion', *Mathematical Problems in Engineering*, Vol. 2020, p.8269683, DOI: 10.1155/2020/8269683.

Xue, J., Li, Z.Y., Wang, X. and Ji, Y.L. (2022) 'Dynamic evaluation and spatial characteristics of smart manufacturing capability in China', *Sustainability*, Vol. 14, No. 17, p.10733, DOI: 10.3390/su141710733.

Yin, H.C., Zhang, G.Z., Wu, Q., Cui, F.P., Yan, B.C., Yin, S.X., et al. (2024) 'Unraveling overlying rock fracturing evolvement for mining water inflow channel prediction: a spatiotemporal analysis using ConvLSTM image reconstruction', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, p.4510417, DOI: 10.1109/TGRS.2024.3452937.

Zhang, Y.P., Nie, S., Liang, S. and Liu, W.J. (2021) 'Robust text image recognition via adversarial sequence-to-sequence domain adaptation', *IEEE Transactions on Image Processing*, Vol. 30, pp.3922–3933, DOI: 10.1109/TIP.2021.3066903.