



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Cross-platform adult learning behaviour profiling based on multimodal data fusion

Na Tian

DOI: [10.1504/IJICT.2026.10076744](https://doi.org/10.1504/IJICT.2026.10076744)

Article History:

Received:	17 November 2025
Last revised:	15 December 2025
Accepted:	17 December 2025
Published online:	17 March 2026

Cross-platform adult learning behaviour profiling based on multimodal data fusion

Na Tian

Directly Affiliated College,
Beijing Open University,
Beijing 100081, China
Email: tiann@bjou.edu.cn

Abstract: In the current era of rapidly developing digital education, single-modal data is insufficient to accurately characterise student learning behaviour. Cross-platform data contains rich information but faces challenges such as data silos and modal heterogeneity. To address this, this paper first collects multi-modal learning behaviour data from students, employs deep learning models to extract text and visual features, and uses attention to fuse multi-modal data. The multimodal fused data is then carefully segmented and distributed to different platform stakeholders. These stakeholders interact and aggregate information through a federated learning framework to obtain more accurate representations of learning behaviour features. Based on this, a Softmax classifier is used to precisely classify cross-platform student learning behaviour profiles. Experimental results show that the proposed method achieves an MAE of only 0.0631, significantly improving the classification accuracy of learning behaviour profiles.

Keywords: cross-platform learning; behavioural profiling; multimodal data fusion; federated learning; deep learning.

Reference to this paper should be made as follows: Tian, N. (2026) 'Cross-platform adult learning behaviour profiling based on multimodal data fusion', *Int. J. Information and Communication Technology*, Vol. 27, No. 22, pp.35–53.

Biographical notes: Na Tian is an assistant researcher in the Directly Affiliated College at Beijing Open University, China. She obtained her Bachelor's in Management (2004) and Bachelor's in Law (2006) from the China University of Political Science and Law, and then obtained her Master's in Public Administration from the Tsinghua University, Beijing, China, in 2013. She has published five papers. Her research interests include adult education, online education and artificial intelligence empowerment education.

1 Introduction

With the deep advancement of digital transformation in education, online learning has become an important part of the education system, and learners' cross-scenario learning behaviours across different platforms are becoming increasingly common (Lee and Roger, 2023). From video-on-demand courses on MOOC platforms to practice feedback from intelligent question banks, learners' learning trajectories are distributed among

diverse carriers, forming fragmented and heterogeneous behaviour data (Shou et al., 2023). These data cover various modalities such as text and images, including explicit records of learning behaviours while also implicitly containing deep information about the learners' cognitive states and learning needs (Zhang, 2022). However, current single-platform learning analytics often fail to comprehensively capture learners' cross-platform learning patterns, leading to an information silo phenomenon in constructed behaviour profiles (Snyman and Van den Berg, 2018). Cross-platform learner behaviour profiling based on multimodal data fusion aims to integrate heterogeneous data dispersed across different learning platforms using artificial intelligence technologies. This process transforms multi-source information into a unified feature representation and ultimately generates comprehensive learner behaviour profiles that reflect the learners' knowledge mastery levels and learning states (Priyambada et al., 2021). This research has significant theoretical and practical implications for promoting educational equity and improving quality.

Student learning behaviour profiling research is an important branch of educational big data and has attracted widespread attention in recent years. Researchers collect and analyse large amounts of student data and use deep learning and traditional machine learning algorithms to build models with the aim of improving the learning experience and providing personalised educational support. Ramírez-Correa et al. (2021) used smart card data and academic performance data to create accurate profiles of students. They used the K-means algorithm (Hu et al., 2023) for clustering and analysed the relationship between academic performance and consumption behaviour. They validated the clustering results using the Apriori algorithm (Jha and Ragha, 2013), generated student tags, and constructed student group profiles using word clouds. Orsoni et al. (2023) used machine learning to construct learning behaviour profiles, based on an improved K-means algorithm using the canopy and maximum minimum distance principles, for student data clustering analysis, and used word clouds to construct student profiles. Song et al. (2024) implemented a college student profiling system, based on research into student profiling theory and technology, optimising the fuzzy C-means algorithm and simulated annealing algorithm to achieve high clustering accuracy. Liang et al. (2021) studied big data platforms, Hadoop, and Hive applications to analyse massive amounts of data to extract student tags, construct student learning behaviour profiles, and optimise tag classification and rules. Tang and Bao (2024) analysed student learning behaviour data generated on the network and used clustering methods to classify users, thereby characterising group user profiles. Students' learning behaviours are dynamic and change over time. As time passes and learning progresses, students' learning styles, interests, and preferences may change. However, traditional machine learning models are typically static. Once trained, they are difficult to adapt to dynamic changes in student learning behaviours, which affects the accuracy and practicality of student profiles.

As the deep learning technique growing, more ideas have been provided for constructing learning behaviour profiles. Deep learning models are highly adaptable and can be applied to different learning scenarios, such as online learning, classroom teaching, and experimental teaching (Kanchon et al., 2024). Whether in traditional school education environments or emerging online education platforms, deep learning models can be adjusted and optimised according to specific learning scenarios and data characteristics to construct student learning behaviour profiles that meet actual needs. Firat (2022) designed and implemented a hybrid neural network model to mine college students' learning behaviour data, establish student profiles, assist students' academic

development, and improve the quality of university teaching. Rogiers et al. (2019) constructed multiple classifiers based on the BERT model (Talaat, 2023) and set different classification weights for different classifiers to classify different forms of text data from students, which was used to construct student learning behaviour profiles and improve the accuracy of the profiles. Kayed et al. (2024) used the latent Dirichlet allocation (LDA) topic model (Yu and Xiang, 2023) and word vector model to mine the meaning of words in students' historical search data, extract learning behaviour features, and construct learning behaviour profiles through word vector representation using long short-term memory network (LSTM) for classification. Students generate a wide variety of data during the learning process, including multimodal learning behaviour data such as text and images. Learning behaviour profiling methods based on a single modality do not fully consider multimodal learning behaviour data, resulting in insufficient feature extraction. Multimodal feature fusion methods provide a solution to the above limitations (Ji et al., 2025). Aydoğdu (2021) used TextCNN and LSTM to extract data such as students' answers and classroom images, respectively, and obtained student profile classification results through a fully connected network.

However, the above study only considered learning behaviour data from a single platform. In reality, data quality varies greatly between different types of data on a single platform, and the total amount of data is relatively small, making it impossible to fully reflect the overall data distribution of the entire online learning platform. Wang and Song (2024) designed a multi-scale convolutional neural network (CNN) to extract features from multimodal learning data and distributed the multimodal fusion data to cross-platform participants through the FL algorithm, reducing data distribution differences between participants and thereby improving the performance of the global model. Shen (2025) constructed a multimodal fusion learning behaviour profile based on multi-platform data. By improving the K-means clustering algorithm, multimodal data was fused at three levels: data, features, and decisions. The FL algorithm was then used to aggregate each layer of the local models of the participating parties with fine-grained attention weights, thereby making the global model closer to the local models in the parameter space.

Based on a comprehensive analysis of existing research on cross-platform learning behaviour profile construction, it is evident that students' learning behaviour data exhibit characteristics of multimodal and cross-platform distribution; however, data silo phenomena across platforms and modality heterogeneity issues lead to shortcomings such as partiality and fragmentation in learning behaviour profiles. To address this, this paper proposes a method for constructing cross-platform adult learning behaviour profiles based on multimodal data fusion. First, students' multimodal learning behaviour data are collected from multiple online learning platforms; the transformer model is used to extract textual semantic features while jointly employing the faster RCNN model to extract visual features, achieving feature fusion of multimodal learning behaviour data through attention. Then, the multimodal fusion data were carefully divided and assigned to different platform participants, each holding their respective datasets. By collaborating jointly with multiple participants in this way, the learning behaviour features embedded in the training dataset can be deeply mined. Each party performs initial feature extraction and analysis using its own local dataset, followed by information exchange and aggregation through an FL framework, thereby obtaining a more comprehensive and accurate representation of learning behaviour characteristics. Based on this, the Softmax classification method is applied to accurately classify students across platforms.

According to the classification results, targeted group-level learning behaviour profiles are constructed to provide strong support for personalised teaching. Experimental results show that the mean absolute error (MAE) of the proposed method is only 0.0631, significantly improving the classification accuracy of learning behaviour profiles. The experimental results demonstrate that the accuracy and F1 scores of the learning behaviour profiles constructed by the proposed method reach 93.72% and 95.39%, respectively, outperforming comparison methods, which fully verifies the efficiency of the proposed approach.

2 Relevant theory

2.1 Federated learning algorithm

FL is a special type of distributed learning (Huang et al., 2024), which includes a central server and N participants $C_1, C_2, \dots, C_N$. FL mainly coordinates multiple participants with different data distributions in relatively complex network environments, aiming to jointly train a more powerful machine learning or deep learning model while protecting data privacy as much as possible and avoiding large-scale data exchange and sharing (Li et al., 2020). In practical applications, FL can not only reduce the storage and computational requirements for training devices caused by large-scale training data in centralised learning but also improve the performance of models trained based on local data, thus ensuring user privacy security for sensitive data.

FL distributes the tasks required for machine learning among multiple clients through collaborative work between a central server and these clients. After each client completes partial model training locally, it uploads the results to the central server; finally, a global model is aggregated at the central server and then returned to the participating clients (Zeng et al., 2023). The objective function of FL algorithms is as follows.

$$f(\omega) = \sum_{k=1}^m \frac{n_k}{n} F_k(\omega) \quad (1)$$

In this equation, m represents the number of participants, n denotes the total amount of data across all participants, n_k indicates the amount of data held by the k^{th} participant, and the local objective function for the k^{th} participant is denoted as $F_k(\omega)$.

$$F_k(\omega) = \frac{1}{n_k} \sum_{i \in P_k} f_i(\omega) \quad (2)$$

In this context, P_k represents a training sample in the k^{th} dataset; the number of samples is indicated by $n_k = |P_k|$, while $f_i(\omega)$ denotes the local loss function.

2.2 Convolutional neural network

CNN is one of the most popular artificial neural network models and has wide applications in computer vision, such as object detection, pose estimation (Luo et al., 2018). Traditional fully connected neural networks require a large number of neurons to extract image features and only accept one-dimensional inputs, which are very unfriendly

for visual tasks. CNN uses convolutional kernels with few neurons to extract image features through smooth movement, reducing the number of model parameters. With the continuous development of computing power and gradually increasing demand for accuracy, deep neural networks have gradually become mainstream. CNN performs matrix calculations using multiple convolutional kernels across multiple channels, achieving nonlinear operations with appropriate activation functions, and finally extracts image depth features through several convolutions (Vonder Haar et al., 2023).

The core components of CNN are the convolutional layer and pooling layer, where each network layer contains multiple two-dimensional planes composed of neurons (Liu et al., 2016). The convolution operation is shown in equation (3), where $*$ represents the convolution operation, w_i denotes the weight value of the i^{th} convolution kernel, b_i indicates the bias term for the i^{th} convolution kernel, $f(x)$ represents the activation function, and L_{i-1} refers to the feature map from the $(i-1)^{\text{th}}$ layer.

$$L_i = f(L_{i-1} * w_i + b_i) \quad (3)$$

After the convolution operation, to reduce the dimensionality of the obtained feature map for improving computational speed and reducing overfitting, convolutional layers and pooling layers are usually connected together as a feature extraction unit. CNN adjusts weights w_i and bias terms b_i through feedback during the training process, thus minimising the loss function. A common loss function in CNN is the cross-entropy function, as shown in equation (4), where n represents the number of samples, x denotes the input value, a is the predicted label, and y refers to the actual value.

$$Loss = -\frac{1}{n} \sum_x [y \ln a + (1-y) \ln(1-a)] \quad (4)$$

CNN has significant advantages in processing images; it can intuitively extract features and simplifies the spatial complexity of network training, greatly enhancing feature representation effectiveness (Zhao et al., 2024).

3 Representation of learning behaviour under multimodal data fusion

3.1 Multimodal data acquisition and pre-processing

As mentioned earlier, students' learning behaviour data are multimodal; considering only unimodal text data cannot accurately depict cross-platform adult learning behaviour profiles. This paper starts from multidimensional spatiotemporal information sources and utilises multimodal data to mutually assist and supervise the semantic information of different modalities, obtaining features of high and low levels of multidimensional engagement. At the same time, the fused feature vector is used as an implicit vector for learners in order to improve the accuracy of subsequent learning behaviour profile construction. Based on the MOOC platform (Perifanou and Economides, 2022) and Coursera platform (Ngo et al., 2023), this paper intelligently collects multimodal data and constructs learner's behavioural features from four dimensions: participation behaviour, social network, cognition, and emotion.

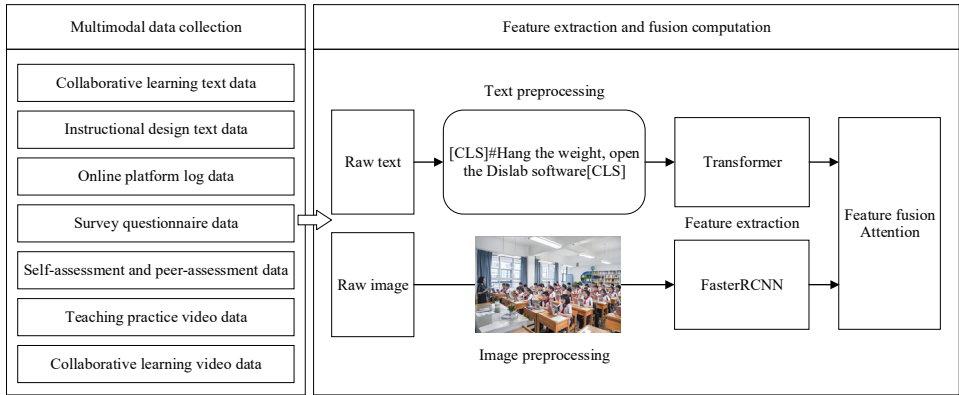
The Scrapy + Selenium data crawling scheme is adopted to obtain static and dynamic page data from the MOOC platform and Coursera platform, parsing the target elements.

Among them, text resources mainly include collaborative learning texts, instructional design texts, online platform logs, questionnaires, self-assessment, and peer assessment texts; video resources include instructional practice videos and collaborative learning videos. After crawling the data, data cleaning is conducted to improve the quality of the data. First, duplicate data, noise data, and dirty data are removed; then, missing data are supplemented and filled in; finally, resource attributes are merged according to the resource ID.

As instructional design texts have relatively professional semantic expression while video audio information has more colloquial semantic expression, visual data can be used for supplementary alignment of multimodal data and characterisation of video content. For this purpose, pre-processing of video data mainly includes three aspects.

- 1 According to the instructional design text, the video is divided into semantically relatively complete segments.
- 2 Extract key frames from video segments that best represent the characteristics of each segment.
- 3 Process the segmented video clips to form a caption text. After data crawling, cleaning and pre-processing, semantically structured data will be stored as the basis for learning behaviour feature extraction.

Figure 1 The multi-modal learning behaviour data feature extraction process (see online version for colours)



3.2 Multi-modal learning behaviour data feature extraction based on deep learning

The learning behaviour representation model under multimodal data fusion is mainly composed of a text feature extractor based on transformer (Chen et al., 2002), an image feature extractor based on faster R-CNN (Ren et al., 2015), and a multimodal fusion component. Specifically, this paper uses word-level pre-trained transformer to extract textual semantic features and generates hidden layer feature vectors. It adopts faster R-CNN to detect targets in videos related to information-based teaching behaviours; learning behaviour feature vectors are then obtained through the multimodal fusion

component. The multi-modal learning behaviour data feature extraction process is shown in Figure 1.

- 1 Transformer text feature extractor. Since learning text data consists of short texts, this paper uses transformer for feature extraction. For example, given a sequence $X = \{\omega_1, \omega_2, \dots, \omega_i\}$ in an instructional design text, where ω_i represents the words in the input sequence and $i \in [1, n]$, the specific modelling method is as follows.

$$e_i = W_e \omega_i \quad (5)$$

$$h_i = \text{Attention}(e_i + p_i) \quad (6)$$

After the model input ω_i is processed by an embedding layer W_e , it is converted into low-dimensional dense feature vectors $e_i \in R^{d_{model}}$ with a vector dimension of d_{model} . Positional encoding consistent with the original transformer paper is adopted, adding the initial word embeddings e_i and positional embeddings p_i . The text information h_i is then explored through multi-head attention (Chen et al., 2020). To fully extract multi-dimensional learning engagement features from textual data, this study's training objective for learning text encoding aims to minimise network loss functions. Input text information h_i goes through average pooling before predicting the categories of learning engagement labels. 15% of the input text information is randomly masked and replaced with a mask token, enhancing semantic representation ability by predicting missing elements.

$$Loss_1 = \text{Softmax}\left(\text{mean}_{pooling}\left(\sum_{i \in \{1, 2, \dots, n\}} h_i\right)\right) \quad (7)$$

$$Loss_2 = \text{Mask}\left(\text{Linear}(h_i)\right) \quad (8)$$

$$Loss = Loss_1 + Loss_2 \quad (9)$$

- 2 Faster RCNN image feature extractor. Instructional videos are composed of images of arbitrary size $P \times Q$. After keyframe images are resized to a fixed size $M \times N$ and input into the feature extraction network, candidate region feature maps are obtained through POI pooling. Fully connected layers map them to a fixed dimension as shown below, where F_v represents fully connected layers, v_j denotes convolutional features of faster R-CNN, and l_j refers to the features after fixed-dimension mapping.

$$l_j = F_v(v_j) \quad (10)$$

3.3 Multimodal feature fusion based on attention mechanism

Textual features h_i are fused with image features l_j , exploring associations between multimodal elements. The relevance $S_{i,j}$ between corresponding words in the learning text and corresponding regions in images is calculated, where h_i^T represents the transpose of h_i .

$$S_{i,j} = h_i^T l_j \quad (11)$$

Since there is an inherent relationship between the text and keyframe images in learning activities, aligning textual semantic information with visual information of keyframe images, as shown in equation (12) and equation (13), where h'_i represents visual information representation for words, and l'_j represents semantic information representation for keyframe images.

$$h'_i = \sum_{j=1}^k \frac{\exp(s_{ij})}{\sum_{m=1}^k S_{im}} l_j \quad (12)$$

$$l'_j = \sum_{i=1}^n \frac{\exp(s_{ij})}{\sum_{m=1}^n S_{mj}} h_i \quad (13)$$

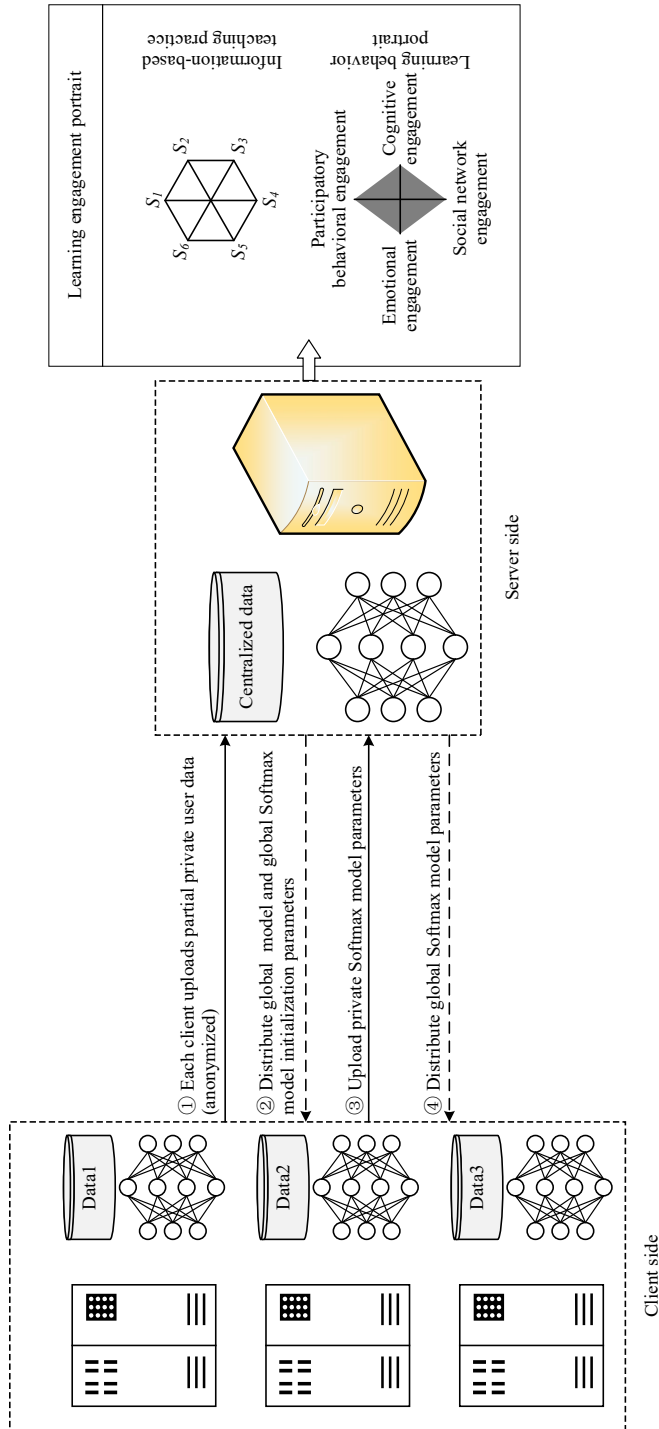
To ensure the validity of multiplying values of different dimensions, individual components with the same operations in textual modality and video modality use Softmax to normalise attention scores, thereby obtaining fused features F .

4 Cross-platform learning behaviour profiling based on multimodal data fusion and federated learning

4.1 Overall framework for building cross-platform learning behaviour profiles

Multimodal data are often distributed across multiple online learning platforms; how to efficiently utilise these multimodal data to precisely construct cross-platform learning behaviour profiles is a challenge (Zhang et al., 2024). Due to large differences in student learning behaviours and uneven distribution of study domains among different online learning platforms, certain user-scarce learning domains may have unclear classifications or even fail to extract their learning behaviours. To address the aforementioned issues, this chapter proposes an adult learning behaviour profiling algorithm based on multimodal data fusion and FL (MFFL), targeting problems such as insufficient cross-platform feature extraction for learning behaviours and inaccurate profiling caused by platform data heterogeneity. The overall flow of the MFFL approach is shown in Figure 2. By employing an architectural pattern based on FL algorithms, deep semantic mining is conducted on multimodal fusion data. Through this process, more comprehensive and rich semantic information can be obtained. At the same time, leveraging the characteristics of FL, multiple online learning platform participants collaborate to achieve precise classification of student learning behaviour profiles. This method effectively resolves many practical issues. On one hand, it overcomes the inaccuracy problems caused by limited data dimensions during the construction of learning behaviour profiles by a single data holder. On the other hand, it breaks down barriers between multiple data holders caused by factors like device differences, avoiding the dilemma of inability to centrally store and share data, thereby solving the resulting challenges in constructing learning behaviour profiles. Compared to the learning behaviour profiles constructed by a single data holder, the profiles built using this method can more comprehensively describe learning behaviour characteristics and achieve more accurate learning behaviour classification.

Figure 2 Process for constructing learning behaviour profiles based on multimodal data fusion and FL (see online version for colours)



In this study, an innovative pseudo-distributed FL architecture is adopted. This architecture takes the multimodal fusion data mentioned earlier as its core training data source. Considering the non-independent and identically distributed nature of data commonly observed in federated learning environments, detailed segmentation was performed on the multimodal fusion data according to specific data types, and it was assigned to different data holders. These holders serve as participants in FL, each holding corresponding datasets. By employing this collaborative approach involving multiple parties, the learning behaviour features embedded in training datasets can be deeply mined. Specifically, each party conducts preliminary feature extraction and analysis using its own dataset locally, followed by information exchange and aggregation through the FL framework to obtain more comprehensive and accurate representations of learning behaviour features. Based on this foundation, a carefully designed FL classification method is applied for precise classification of cross-platform student behaviour profiles. Ultimately, representative group-level learning behaviour profiles with specificity are constructed based on the classification results, providing strong support for subsequent educational research and instructional decision-making.

Firstly, the distribution of student types among participants on each learning platform is statistically analysed, with categories having higher proportions defined as major categories and those with lower proportions defined as minor categories. Subsequently, a portion of local data from the major category students in each participant's dataset is extracted to serve as training datasets for the centre server to train the global model and uploaded to the central server. The central server pre-processes the collected partial student data from each participant and trains the global model based on this dataset to mine behavioural features across all participants. After completing the model training, it is distributed back to each participant and used as a pre-processing model for subsequent learning behaviour profile classification tasks. The globally trained model by the central server can effectively address issues such as inaccurate classifications caused by imbalanced data distribution or poor data quality at individual participants' ends.

Within the FL collaborative framework, after each participant receives the global model from the central server, they use this model to perform in-depth pre-processing on their locally held data. Through this pre-processing stage, it is possible to accurately calculate probability weights for each student across various predefined topics; these probability weights reflect students' tendencies and correlations toward different learning topics. Subsequently, each participant utilises the Softmax classification algorithm (Wang et al., 2021) to conduct training work for student categorisation based on local data. Once the classification model has been trained, participants can efficiently perform student classification operations on their own data using the results generated by the global classification model. Through this classification process, representative student group learning domain labels can be extracted from massive data. These labels cover feature information regarding students' performance across different subjects and learning activities. Ultimately, based on these learning domain labels, a comprehensive, detailed, and targeted group-level student behaviour profile is constructed. The overall framework of MFFL is shown in Figure 2.

4.2 MFFL central server algorithm

Participants from each online learning platform first extract major category student data from their local private datasets, then select student behavioural data samples with

distinct features to support collaborative multi-modal data topic feature mining across all FL participants. As a trusted third party, the central server is responsible for securely aggregating de-identified student data from all participants and achieving centralised management and analysis under privacy protection constraints, constructing an evenly distributed pre-training dataset $\{\langle \bar{V}_1, k_1 \rangle, \dots, \langle \bar{V}_M, k_K \rangle\}$, where $\bar{V}_1$ represents multi-modal fused student data and k_1 denotes the category labels of students. This dataset adopts a centralised access control mechanism that only allows authorised access by the central server; data sharing and cross-participant access are strictly prohibited. The central server will utilise this pre-training dataset to complete the training tasks for the global model.

The selection of the number of topics during global modelling is crucial. Current research often employs heuristic rules or perplexity metrics to determine the number of topics for multi-modal fused data. This paper proposes a dual-index evaluation system based on perplexity and topic consistency, optimising the choice of the number of topics through a dual-validation mechanism. Perplexity quantifies the uncertainty in model classification results. This study adopts a joint optimisation strategy. Specifically, a lower perplexity value indicates better classification performance by the model. However, when the volume of multi-modal data increases, perplexity decreases mechanically, which may lead to overfitting on training data and thereby weaken the practical interpretability of the resulting model. First, it filters candidate topic quantity ranges through perplexity and then performs precise calibration based on consistency scores. Ultimately, the optimal number of topics is determined by dual-index weighted evaluation to ensure model classification performance while enhancing the interpretability of the topics.

In the federated learning framework, the central server first initialises the parameters of the global Softmax classification model. Subsequently, the system dynamically determines the number of participants n for each round of model updates based on a preset participation ratio and selects n nodes through a random sampling mechanism from all available participants to form the FL participant set for the current training round P_g . The initial parameters are distributed to P_g , and in each round of FL iteration, each participant first retrieves the latest global model parameters from the central server. Then, each node independently performs model training and parameter updates using local private data. Finally, all selected participants must synchronise their locally trained model parameters to the central server for subsequent global model aggregation operations. The central server calculates corresponding aggregation weights based on the proportion of local training data volume provided by each participant relative to the total data volume in this round of global training and executes weighted parameter aggregation. This process iterates until it meets the preset FL training round termination condition.

4.3 MFFL participant-side algorithm

After completing pre-processing on the local private dataset, participants first statistically analyse students' behavioural data volumes across categories, designating categories with a high proportion of data volume as major categories; then they calculate word frequencies in student data and randomly sample some samples from the large-category student data that contain high-frequency information to upload to the central server, providing data support for training the global model.

Relevant students on the platform D_i receive the global model provided by the central server and map local training data into dimensions of the topic feature space, with any sample j expressed as $V_j = \{\langle T_1, q_1 \rangle, \langle T_2, q_2 \rangle, \dots, \langle T_n, q_n \rangle\}$, where n is the number of learning topics and the local training data for participant i is $C_i = \{\langle V_1, k_1 \rangle, \dots, \langle V_{|C_i|}, k_K \rangle\}$.

During the local model training phase, participants first perform topic feature dimension extraction on raw data and then use the resulting feature vectors as input to train a local classification model based on the Softmax function, where the model parameters are $u = (\vartheta_1, \dots, \vartheta_K)$, and the output is the probability of students V_j across categories, denoted as $h_\tau(V_j)$, as shown in equation (14).

$$h_\tau(V_j) = \begin{bmatrix} p(k_j=1|V_j; \tau) \\ \dots \\ p(k_j=K|V_j; \tau) \end{bmatrix} = \frac{1}{\sum_{k=1}^K e^{(\tau * V_j)_k}} \begin{bmatrix} e^{\vartheta_1 V_j} \\ \dots \\ e^{\vartheta_K V_j} \end{bmatrix} \quad (14)$$

where τ is the current local model parameter, and $\vartheta_K V_j$ is the value of the K^{th} category output by the output layer of the model. The probability that the learning behaviour V_j belongs to category k is as shown in equation (15).

$$P(k_j = k | V_j; \tau) = \frac{e^{\vartheta_k V_j}}{\sum_{k=1}^K e^{\vartheta_k V_j}} \quad (15)$$

In multi-classification tasks based on Softmax, the model output satisfies the normalisation constraint for probabilities; that is, for any input sample, the sum of its predicted probabilities across K categories is always 1. Model training employs the cross-entropy loss function as an optimisation objective, which quantifies the discrepancy between the predicted probability distribution and the true label distribution, as shown in equation (16).

$$L(\tau) = -\frac{1}{|C_i|} \left[\sum_{j=1}^{|C_i|} \sum_{k=1}^K 1\{k_j = k\} \log \frac{e^{\vartheta_k V_j}}{\sum_{k=1}^K e^{\vartheta_k V_j}} \right] \quad (16)$$

where $1\{\}$ is an indicator function. The gradient of model parameters ϑ_q can be solved using equation (17).

$$\frac{\partial L(\tau)}{\partial \vartheta_q} = -\frac{1}{|C_i|} \frac{\partial}{\partial \vartheta_q} \left[\sum_{j=1}^{|D_i|} \sum_{k=1}^K 1\{k_j = k\} \log \frac{e^{\vartheta_k V_j}}{\sum_{k=1}^K e^{\vartheta_k V_j}} \right] \quad (17)$$

In the global classification model training phase, D_i receives the current round's global model parameters from the central server and updates the local training model. The participants train the local model parameters τ_e^i according to $L(w)$. After completing the local model training, the participants upload the updated model parameters to the central server for participation in the global model aggregation. The system then enters a standby state, waiting for the trigger of the next round of FL iteration.

Through this training process, the global model can effectively integrate data from multiple participants and deeply mine students' learning behaviour features; further train

cross-platform global classification models based on these features to improve user classification accuracy for individual participants. This allows each participant to more accurately capture student characteristics and promote the construction of learning behaviour portraits toward a more complete and accurate direction.

5 Experimental results and performance analysis

5.1 Learning behaviour portrait output result analysis

The experimental dataset of this paper comes from the MOOC platform and Coursera platform, crawling students' behaviour logs, text data, learning videos, and other data from these two platforms. The datasets from both platforms are pre-processed to extract 36,021 records of learning behaviour data for users who have studied on both platforms. The dataset is divided into three parts: 70% used for model training, 20% for validation, and the remaining 10% for testing the model during training. The experiments are run on a computer equipped with an Intel (R) Core (TM) i7-9700K CPU, 16 GB of memory, and four NVIDIA GeForce RTX 2070 graphics cards. Python 3.7.9 is used as the main programming language, and experiments are conducted on the Ubuntu 20.04 LTS system. In the experiment, SGD learning rate is set to 0.01, batch size is 16, and the number of attention heads in transformer is set to 4.

This paper categorises the learning groups in the dataset into three types for learning behaviour portrait analysis based on a score threshold of 0.5 across various dimensions. Learners in group one have scores greater than 0.5 in all seven dimensions, showing well-balanced performance with an overall learning ability level recorded as A, as shown in Figure 3(a). Learners in group two have a score below 0.5 on the dimension of learning status but above 0.5 for other dimensions; their ability to engage in friendly communication during learning needs further improvement, and their behaviour level is recorded as B, as shown in Figure 3(b). Learners in group three show scores slightly greater than 0.5 only in knowledge mastery and leadership, while the rest are below 0. In the learning process, special attention should be given to developing this type of learner's collaborative abilities, with their behaviour level recorded as C, as shown in Figure 3(c).

Individual learning behaviour portraits can help teachers accurately understand the detailed situation of a specific learner, with an example of student Zhang to present their individual learner portrait and related explanations. As shown in Figure 3(d), an excellent learner is selected from the dataset, whose learning behaviour level is A, academic performance is A, and overall performance is good, showing strong leadership skills and emotional regulation abilities; teachers should focus on cultivating this learner's collaborative intentions.

5.2 Performance analysis of learning behaviour portrait construction

The previous chapter provided a detailed analysis of the learning behaviour portrait output results. This chapter introduces quantitative evaluation metrics accuracy, F1, AUC, and MAE to assess the construction accuracy of different methods. Our method is denoted as OURS, while baseline methods include ConvLSTM (Aydoğdu, 2021), DMSCNN (Wang and Song, 2024), and MFKMS (Shen, 2025). The portrait construction accuracies of different methods are presented in Table 1. OURS achieves an accuracy and

F1 of 93.72% and 95.39%, respectively, showing improvements of 12.54% and 12.3% over ConvLSTM, 8.25% and 11.73% over DMSCNN, and 3.66% and 3.54% over MFKMS. Comparing the area under the ROC curve bounded by AUC, OURS achieves an AUC of 0.9801, showing improvements of 8.21%, 6.96%, and 3.86%. From the construction accuracy metric MAE, it can be seen that MAE has a low classification error for learning behaviour portraits, with an MAE of only 0.0631, significantly improving the classification accuracy of learning behaviour portraits. OURS not only performs feature extraction and fusion on multimodal learning data from multiple platforms but also designs corresponding FL algorithms to design global and local models across multiple platforms. It effectively mines students' learning behaviour features using a global model for multiple participants' data, thereby making the construction of student learning behaviour portraits more complete and accurate.

Figure 3 Group and individual learning behaviour profiles, (a) group one: learner behaviour profile, (b) learning behaviour profile of learners in category II, (c) profile of learners in category III, (d) individual learning behaviour profile (see online version for colours)

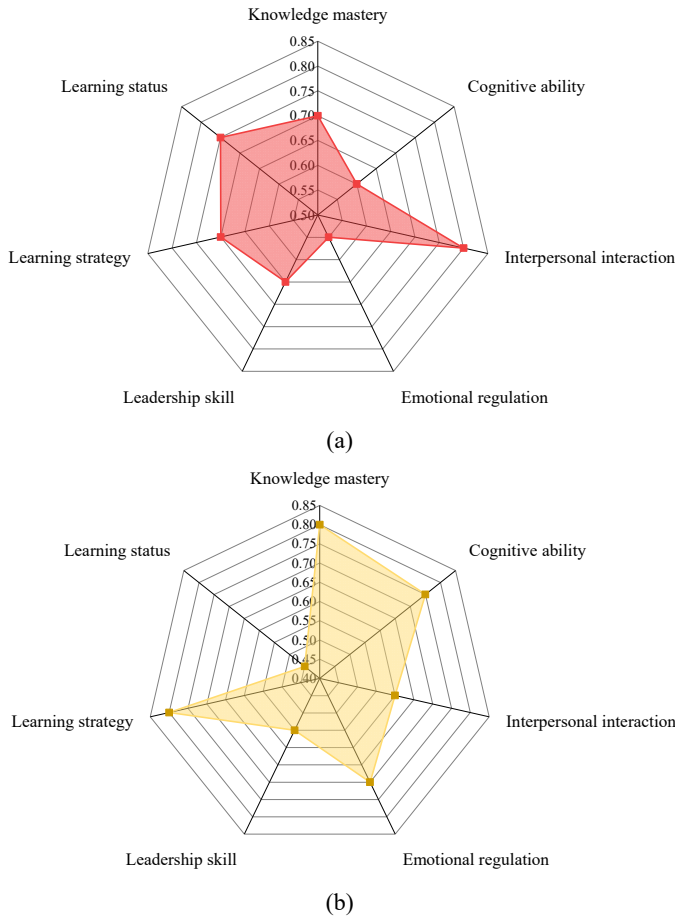


Figure 3 Group and individual learning behaviour profiles, (a) group one: learner behaviour profile, (b) learning behaviour profile of learners in category II, (c) profile of learners in category III, (d) individual learning behaviour profile (continued) (see online version for colours)

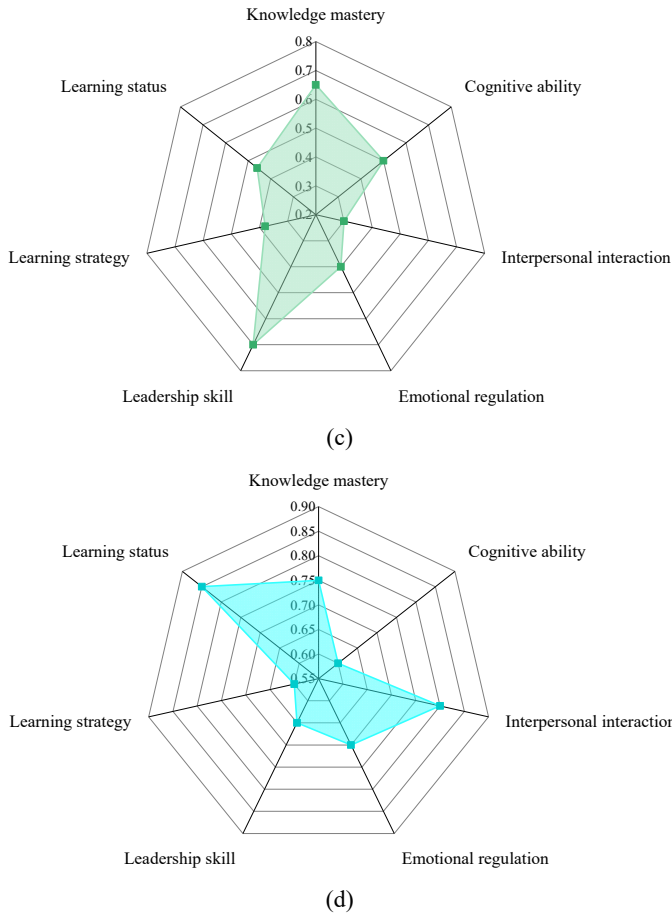


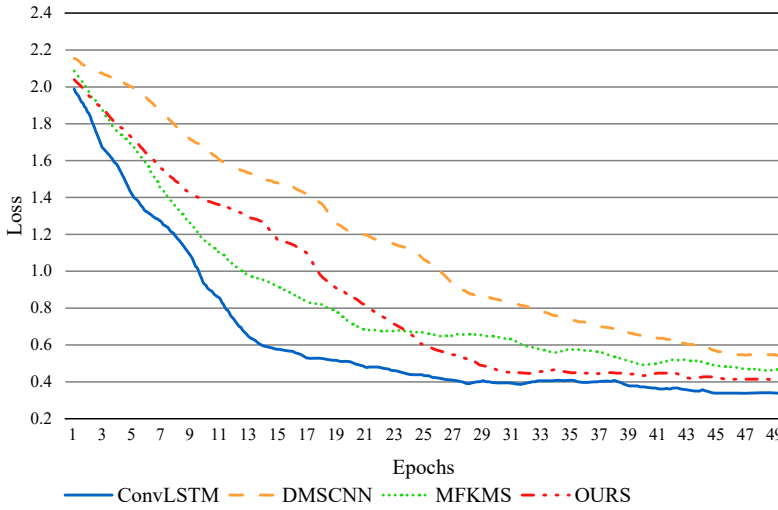
Table 1 Accuracy of learning behaviour profiling using different methods

Method	Accuracy/%	F1/%	AUC	MAE
ConvLSTM	81.18	83.09	0.9057	0.1384
DMSCNN	85.47	83.66	0.9163	0.1297
MFKMS	90.06	91.85	0.9437	0.0631

The convergence speed of the loss curves for each method is compared in Figure 4. OURS can reduce cross-entropy loss to a small value by the 14th iteration, whereas MFKMS has not fully converged even by the 45th iteration. DMSCNN begins to stabilise its cross-entropy loss around the 21st iteration, but it remains generally higher than that of OURS. ConvLSTM converges the slowest; only at the 31st iteration does the loss change for ConvLSTM gradually diminish. The ConvLSTM method simultaneously optimises convolutional parameters for TextCNN and recurrent parameters for LSTM,

which may lead to unstable training or overfitting, possibly preventing the model from fully learning features. The DMSCNN method utilises CNN and FL for constructing cross-platform learning behaviour profiles, but learning behaviour data are usually sparse and irregular, while CNN requires input data to have fixed dimensions; direct padding or truncation may lead to information loss or noise introduction. MFKMS trains K-means through FL for analysing cross-platform learning behaviours, but the model might generate fragmented clustering results due to data heterogeneity, such as the same behaviour being divided into different clusters by different clients, which reduces the generality of profiles. Compared to baseline methods, OURS not only achieves feature extraction for text and images through transformer and faster RCNN, performs multimodal fusion via attention mechanisms, but also iteratively trains the results of multimodal fusion through FL, greatly improving classification accuracy for learning behaviour profiles.

Figure 4 Convergence speed of loss curves for each method (see online version for colours)



6 Conclusions

Against the backdrop of the widespread popularity of digital education, students' learning behaviour data exhibit characteristics of being multimodal and cross-platform distributed; however, data island phenomena among platforms as well as modality heterogeneity issues result in partialised and fragmented defects in learning behaviour profiles. To solve the above problems, this study aims to construct a cross-platform adult learning behaviour profile based on multimodal data fusion for providing scientific basis for personalised teaching. First, we collect students' multimodal learning behaviour data from multiple online learning platforms, employ transformer models to extract textual semantic features and jointly use faster RCNN models to extract visual features, achieving feature fusion of multimodal learning behaviour data through attention. Then, the multimodal fusion data were carefully partitioned and allocated to different platform participants, each holding the corresponding dataset. Each party conducts preliminary

feature extraction and analysis using their local datasets, followed by information exchange and aggregation through an FL framework to obtain a more comprehensive and accurate representation of learning behaviour features. Based on this, each participating party employs the Softmax classification algorithm to train student behaviour profile categorisation using local data, extracting representative learning domain labels from massive data for student groups. Finally, based on these learning domain labels, a comprehensive, detailed, and targeted group-level student behaviour profile is constructed. Experimental results show that the construction accuracy of the proposed method reaches 93.72%, showing significant improvement over baseline methods, enabling the construction of relatively complete and accurate cross-platform learning behaviour profiles.

With the deep integration of information-based teaching and disciplinary knowledge, the next phase will primarily focus on cognitive graphs based on learning engagement to assist in constructing cross-platform learning behaviour profiles. Attributes such as achievement status of learning objectives, learning preferences, and cognitive levels will be mined from learning engagement features. For unmet learning objectives, attribution analysis using a cognitive graph can effectively support profile construction and enhance the efficiency of personalised teaching.

Declarations

All authors declare that they have no conflicts of interest.

References

- Aydoğdu, Ş. (2021) ‘A new student modeling technique with convolutional neural networks: learnerprints’, *Journal of Educational Computing Research*, Vol. 59, No. 4, pp.603–619.
- Chen, H., Jiang, D. and Sahli, H. (2020) ‘Transformer encoder with multi-modal multi-head attention for continuous affect recognition’, *IEEE Transactions on Multimedia*, Vol. 23, pp.4171–4183.
- Chen, T-H., Chen, M-S., Inoue, T., Kotas, P. and Chebli, E.A. (2002) ‘Three-phase cogenerator and transformer models for distribution system analysis’, *IEEE Transactions on Power Delivery*, Vol. 6, No. 4, pp.1671–1681.
- Firat, M. (2022) ‘Profiling learning preferences of distance education students based on neural network analysis’, *Turkish Online Journal of Distance Education*, Vol. 23, No. 2, pp.1–14.
- Hu, H., Liu, J., Zhang, X. and Fang, M. (2023) ‘An effective and adaptable K-means algorithm for big data cluster analysis’, *Pattern Recognition*, Vol. 139, pp.109–124.
- Huang, W., Wang, D., Ouyang, X., Wan, J., Liu, J. and Li, T. (2024) ‘Multimodal federated learning: concept, methods, applications and future directions’, *Information Fusion*, Vol. 112, pp.82–97.
- Jha, J. and Ragha, L. (2013) ‘Educational data mining using improved Apriori algorithm’, *International Journal of Information and Computation Technology*, Vol. 3, No. 5, pp.411–418.
- Ji, X., Sun, L. and Huang, K. (2025) ‘The construction and implementation direction of personalized learning model based on multimodal data fusion in the context of intelligent education’, *Cognitive Systems Research*, Vol. 4, pp.101–119.

- Kanchon, M.K.H., Sadman, M., Nabila, K.F., Tarannum, R. and Khan, R. (2024) 'Enhancing personalized learning: AI-driven identification of learning styles and content modification strategies', *International Journal of Cognitive Computing in Engineering*, Vol. 5, pp.269–278.
- Kayed, M., Azzam, F., Ali, H. and Ali, A. (2024) 'Temporal dynamics of user activities: deep learning strategies and mathematical modeling for long-term and short-term profiling', *Scientific Reports*, Vol. 14, No. 1, pp.14–25.
- Lee, Y-J. and Roger, P. (2023) 'Cross-platform language learning: a spatial perspective on narratives of language learning across digital platforms', *System*, Vol. 118, pp.103–115.
- Li, L., Fan, Y., Tse, M. and Lin, K-Y. (2020) 'A review of applications in federated learning', *Computers & Industrial Engineering*, Vol. 149, pp.82–104.
- Liang, K., Liu, J. and Zhang, Y. (2021) 'The effects of non-directional online behavior on students' learning performance: a user profile based analysis method', *Future Internet*, Vol. 13, No. 8, pp.99–113.
- Liu, M., Shi, J., Li, Z., Li, C., Zhu, J. and Liu, S. (2016) 'Towards better analysis of deep convolutional neural networks', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 23, No. 1, pp.91–100.
- Luo, H., Xiong, C., Fang, W., Love, P.E., Zhang, B. and Ouyang, X. (2018) 'Convolutional neural networks: Computer vision-based workforce activity assessment in construction', *Automation in Construction*, Vol. 94, pp.282–289.
- Ngo, T.T.A., Tran, T.T., An, G.K. and Nguyen, P.T. (2023) 'Students' perception towards learning massive open online courses on Coursera platform: benefits and barriers', *International Journal of Emerging Technologies in Learning*, Vol. 18, No. 14, pp.70–82.
- Orsoni, M., Giovagnoli, S., Garofalo, S., Magri, S., Benvenuti, M., Mazzoni, E. and Benassi, M. (2023) 'Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile', *Heliyon*, Vol. 9, No. 3, pp.83–96.
- Perifanou, M. and Economides, A.A. (2022) 'The landscape of MOOC platforms worldwide', *International Review of Research in Open and Distributed Learning*, Vol. 23, No. 3, pp.104–133.
- Priyambada, S.A., Er, M., Yahya, B.N. and Usagawa, T. (2021) 'Profile-based cluster evolution analysis: identification of migration patterns for understanding student learning behavior', *IEEE Access*, Vol. 9, pp.101718–101728.
- Ramírez-Correa, P., Alfaro-Pérez, J. and Gallardo, M. (2021) 'Identifying engineering undergraduates' learning style profiles using machine learning techniques', *Applied Sciences*, Vol. 11, No. 22, pp.105–128.
- Ren, S., He, K., Girshick, R. and Sun, J. (2015) 'Faster R-CNN: towards real-time object detection with region proposal networks', *Advances in Neural Information Processing Systems*, Vol. 28, pp.71–84.
- Rogiers, A., Merchie, E. and Van Keer, H. (2019) 'Learner profiles in secondary education: occurrence and relationship with performance and student characteristics', *The Journal of Educational Research*, Vol. 112, No. 3, pp.385–396.
- Shen, Y. (2025) 'Building an intelligent education model for student profiling based on big data algorithms', *Journal of Intelligent Computing Volume*, Vol. 16, No. 1, pp.65–73.
- Shou, Z., Shi, Z., Wen, H., Liu, J. and Zhang, H. (2023) 'Learning peer recommendation based on weighted heterogeneous information networks on online learning platforms', *Electronics*, Vol. 12, No. 9, pp.20–33.
- Snyman, M. and Van den Berg, G. (2018) 'The significance of the learner profile in recognition of prior learning', *Adult Education Quarterly*, Vol. 68, No. 1, pp.24–40.
- Song, R., Pang, F., Jiang, H. and Zhu, H. (2024) 'A machine learning based method for constructing group profiles of university students', *Heliyon*, Vol. 10, No. 7, pp.51–63.
- Talaat, A.S. (2023) 'Sentiment analysis classification system using hybrid BERT models', *Journal of Big Data*, Vol. 10, No. 1, pp.110–125.

- Tang, H. and Bao, Y. (2024) 'Self-regulated learner profiles in MOOCs: a cluster analysis based on the item response theory', *Interactive Learning Environments*, Vol. 32, No. 5, pp.1877–1893.
- Vonder Haar, L., Elvira, T. and Ochoa, O. (2023) 'An analysis of explainability methods for convolutional neural networks', *Engineering Applications of Artificial Intelligence*, Vol. 117, pp.105–126.
- Wang, H. and Song, Y. (2024) 'Portrait of college students' online learning behavior based on artificial intelligence technology', *IEEE Access*, Vol. 12, pp.6318–6328.
- Wang, Y., Gui, G., Gacanin, H., Adebisi, B., Sari, H. and Adachi, F. (2021) 'Federated learning for automatic modulation classification under class imbalance and varying noise condition', *IEEE Transactions on Cognitive Communications and Networking*, Vol. 8, No. 1, pp.86–96.
- Yu, D. and Xiang, B. (2023) 'Discovering topics and trends in the field of artificial intelligence: using LDA topic modeling', *Expert Systems with Applications*, Vol. 225, pp.114–123.
- Zeng, D., Liang, S., Hu, X., Wang, H. and Xu, Z. (2023) 'Fedlab: a flexible federated learning framework', *Journal of Machine Learning Research*, Vol. 24, No. 100, pp.1–7.
- Zhang, K., Xing, S. and Chen, Y. (2024) 'Research on cross-platform digital advertising user behavior analysis framework based on federated learning', *Artificial Intelligence and Machine Learning Review*, Vol. 5, No. 3, pp.41–54.
- Zhang, R. (2022) 'A personalized course resource recommendation method based on deep learning in an online multi-modal multimedia education cloud platform', *International Journal of Information Technologies and Systems Approach (IJITSA)*, Vol. 16, No. 2, pp.1–14.
- Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M. and Parmar, M. (2024) 'A review of convolutional neural networks in computer vision', *Artificial Intelligence Review*, Vol. 57, No. 4, pp.26–38.