Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **11**, Number **1**, February **2026**

Personalized Learning Path Recommendation for Preschool Education Based on Multimodal Learning and Data Analysis

Liang-Liang Su[1,*]

[1]Yinchuan University of Science and Technology of Education,
YinChuan 750021, P. R. China
1341219230@qq.com

Xiao Sun[2]

[2]Gomel State University,
Gome 246003, Belarus
215110644@qq.com

[*]Corresponding author: Liang-Liang Su

Received January 14, 2025, revised May 09, 2025, accepted August 30, 2025.

Abstract. *How to employ current technology to customise appropriate learning routes for every kid has become a research hotspot with the growing need for customised learning in preschool education. Based on multimodal learning and data analysis, this paper presents a personalised learning route suggestion method for preschool education. First, extract multidimensional features using deep learning models by means of multimodal learning techniques, which combine input from several senses including visual, aural, and motion perception. Convolutional neural networks (CNN) and paired with Long Short Term Memory Networks (LSTM) for time series analysis capture children's learning progress and emotional changes by feature extraction of this multimodal data. Second, learning behaviour patterns and cluster children's performance in various learning activities using K-means technology. Personalised learning paths are suggested for children depending on their learning preferences and progress by use of collaborative filtering systems. The suggested approach can help maximise the teaching efficacy of preschool education in practice and efficiently increase the accuracy of personalised learning recommendations, so supporting the rational distribution of educational resources and personal development.*
Keywords: multimodal; data analysis; personalized learning path recommendation.

1. **Introduction.** Personalised education has grown to be one of the main areas of research in the field of education as early education gets more and more of importance in society. Apart from being a vital component of a child's developmental process, school education serves as the fundamental stage for the acquisition of their cognitive, emotional, and social skills. By now every child's cognitive level, interests, learning style, emotional condition, etc. differ greatly. Thus, in present preschool education, designing individualised learning routes for every kid has become a difficulty. Recommendations for personalised learning paths can offer customised learning materials and teaching strategies depending on the traits of the students, thereby fostering their self-learning capacity and creative thinking as well as aid to increase the effectiveness of education.

Personalised education recommendation systems have been extensively investigated and implemented in recent years with the fast growth of artificial intelligence technology, particularly the maturity of deep learning and data analysis techniques. Research on personalised learning route recommendations has also drawn more and more importance in the field of preschool education [1]. Usually emphasising big class instruction, the conventional educational paradigm ignores the particular needs of every student. Modern technology, particularly the adoption of multimodal learning strategies, enables teachers to more precisely depict the several ways in which students participate in the learning process and offer tailored learning recommendations using big data analysis [2].

Emerging in recent years, multimodal learning is a new research paradigm distinguished by the building of more complete learning models by use of information from several sensory channels [3]. In preschool education, multimodal data refers to children's visual, aural, and physiological feedback—that is, emotional responses, attention states, etc.—that transcends their visual information. Rich contextual support for learning path recommendations can come from the interactions of this information. Children's learning demands and emotional changes are sometimes not completely reflected in traditional single data sources (like linguistic data or behavioural data). Thus, employing multimodal data can more fully assess children's learning situation, so obtaining more accurate personalised learning path recommendations.

Preschool children's learning behaviour is highly diverse and unpredictable, which presents several difficulties for recommendation systems in use [4]. First of all, data processing and analysis find great challenges in preschool education since much of the data is unorganised and varied. Second, there are great individual variations in the learning development and emotional state of preschoolers; so, conventional recommendation systems sometimes find it difficult to meet these particular needs. Thus, a major challenge in present research is how to efficiently extract significant features from huge multimodal data and mix these characteristics with learning path recommendation models.

This work suggests a customised learning route recommendation technique based on multimodal learning and data analysis to handle these problems for preschool education. This approach integrates CNN and LSTM models in deep learning to extract individualised learning traits of children from multimodal data including visual, auditory, and action data. Furthermore employed are cluster analysis methods to categorise children's learning patterns, spot several learning style groups, and suggest the best course of action for kids from several backgrounds. Concurrently, the accuracy of learning path recommendations is further enhanced depending on the child's past learning behaviour and preferences by means of cooperative filtering algorithms.

1.1. **Related work.** Many academics are dedicated to investigate personalised learning route recommendation in preschool education as the research on this topic has attracted more attention in recent years especially with the ongoing development of multimodal learning and data analysis technologies. Early studies mostly concentrated on developing personalised learning recommendations using a single data source—text, behavioural, or emotional data. Based on student behaviour data, Kim and Park [5] suggested a customised suggestion system that enables teachers to change their approaches by means of analysis of student behaviour during the course of instruction. Although this approach is efficient, its single data source limits cause the recommendations to be approximative and neglect the multidimensional aspects of pupils. More and more research are trying to mix several data sources to raise the accuracy of recommendation systems as multimodal learning grows. Zhang [6] jointly models the data using CNN and LSTM models to get more extensive knowledge on student learning performance and uses multimodal

data (such as language, vision, and emotion) for personalised learning recommendations. While this approach has made great strides, how best to combine data from several modalities remains a major challenge to be resolved. Tripathi et al. [7] suggested another comparable study whereby sentiment analysis was integrated with multimodal learning approaches based on image and video data to assess students' learning emotions and offer emotional driven personalised learning paths for them. This approach, nonetheless, has not been thoroughly investigated for tailored recommendations of learning content and mostly concentrates on sentiment computing.

How to manage the merging of multimodal data and feature selection in personalised learning recommendation systems has grown to be a major problem. Based on deep neural networks, Ullah et al. [8] introduced a feature fusion technique whereby an adaptive learning mechanism automatically chooses the most relevant features to increase the efficacy of the learning path recommendation. Although this approach has shown good performance in practice, it still presents difficulties reaching effective training in vast data sets. Many studies have also started to concentrate on customised recommendation strategies grounded on cluster analysis. Combining K-means clustering with collaborative filtering techniques, Bhaskaran and Santhi [9] suggested a hybrid recommendation system that divides students into several groups depending on their learning behaviour characteristics and suggests the most appropriate learning path for every group. Although group segmentation still has room for development, this approach can help to considerably raise the accuracy of tailored recommendations.

In the particular context of preschool education, Crescenzi-Lanna [10] investigated how to suggest suitable learning activities for preschoolers using multimodal data and made some development by analysing children's learning behaviour using deep convolutional neural networks (DCNN). Still, it is difficult to combine these approaches with other elements like social contact and emotional development of youngsters. As recommendation systems evolve constantly, many researchers have started to concentrate on ways to increase the real-time and adaptability of recommendations. Based on time-series data, Zhou et al. [11] proposed a customised recommendation system that forecasts children's future learning behaviour using an LSTM model and modulates the learning path depending on the prediction outcomes. Although dynamic personalised recommendations are achieved with this approach, how best to handle learning progress's ambiguity remains a challenge.

Data privacy and security concerns have also drawn lot of attention in the field of preschool education. Using federated learning methods to provide tailored recommendations while guaranteeing user privacy, Xu and Yin [12] suggested a recommendation system framework combining privacy protection features. Although the framework has made considerable development in privacy protection, it is still difficult to balance the conflict between privacy protection and suggestion accuracy. Another relevant area of study is how to improve tailored learning suggestions by means of sentiment analysis Bang et al. [13] maximised learning path suggestions by using children's emotional reactions—that of facial expressions and verbal emotions. Sentiment analysis helps recommendation systems to more precisely offer learning recommendations and better fit to children's emotional fluctuations during the learning process. Lastly, researchers also have great interest in knowing how well tailored learning recommendation systems work. Covering multidimensional variables including accuracy, real-time performance, and user satisfaction, Zheng et al. [14] established a recommendation system assessment framework based on numerous evaluation indicators, therefore offering practical basis for optimising personalised recommendation systems.

1.2. **Contribution.** This work uses multimodal learning approaches, which include multidimensional data from visual, auditory, and motor elements, therefore allowing the recommendation system to have a more complete awareness of children's learning state. Second, important elements can be effectively identified from challenging preschool education data by merging deep learning models with clustering analysis techniques, therefore offering individualised learning recommendations for children; At last, the implementation of collaborative filtering techniques enhances the practicality and correctness of learning route recommendation even more.

2. Theoretical analysis.

2.1. **Convolutional neural network.** Widely applied in disciplines including image identification, object detection, and speech processing, CNN is a deep learning model with strong image processing capability [15]. CNN's design inspiration derives from the perception mechanism in the human visual system, which possesses pooling operation, weight sharing, and local receptive field. Figure 1 displays the CNN model construction.

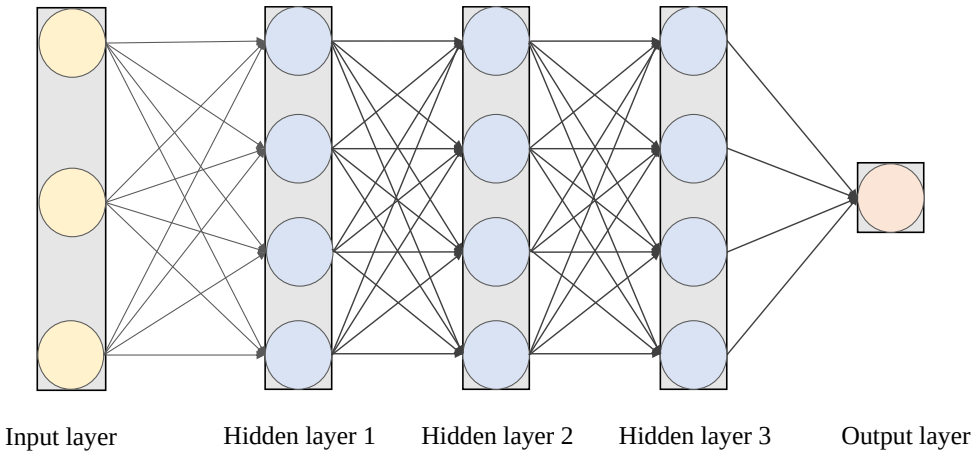

Figure 1. The architecture of CNN

In convolutional neural networks, the most crucial function is convolution, in which sliding window operations on input data via convolution kernels (or filters) extract local characteristics [16]. The convolution operation is defined assuming the input image is a two-dimensional matrix X and the convolution kernel is a small-sized matrix K as:

$$Y(i,j) = (X * K)(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n) K(m,n) \tag{1}$$

where, $Y(i,j)$ represents an element in the convolution output matrix, $X(i,j)$ is the pixel value in the input image, $K(m,n)$ is an element in the convolution kernel, and m and n are the sizes of the convolution kernel. The convolution kernel slides on the input image and performs point by point multiplication and summation to generate the convolution output matrix. Convolution operation can effectively extract local features in images, such as edges, corners, etc.

Convolution operation produces a nonlinear change via an activation function to raise the model's expressive capability. ReLU ("Rectified Linear Unit") is the most often used activation function with an equation:

$$f(x) = \max(0, x) \tag{2}$$

The ReLU activation function accelerates the network's training process and can reasonably prevent gradient vanishing issues. ReLU's graphical depiction is that, should the input be more than zero, the output is the input value; should the input be less than zero, the output is zero. ReLU's benefits are simplicity and great computational economy.

Commonly used downsampling methodologies to lower spatial size of feature maps, lower computational complexity, and avoid overfitting include pooling operations. Usually, the pooling layer makes advantage of either average or max pooling. Consider max pooling: its definition is as follows:

$$Y(i,j) = \max\{X(i+k, j+l) \mid 0 \le k < h, 0 \le l < w\} \tag{3}$$

where, h and w are the height and width of the pooling window. Maximum pooling preserves the most significant characteristics by choosing the maximum value from the local area of the input feature map. By means of translation invariance, effective reduction of feature map size, and computing simplicity, pooling operations can thereby improve the model's robustness.

The Fully Connected Layer (FC) is located at the end of the convolutional neural network and is used to map the features extracted by the convolutional and pooling layers to the final output. The fully connected layer connects all neurons of the previous layer with each neuron of the current layer, thereby achieving the aggregation of global features. Assuming that the input of a certain layer is $x_1, x_2, \ldots, x_n$ and the output of a fully connected layer is $y_1, y_2, \ldots, y_m$, the output value is calculated using the following equation:

$$y_j = \sigma\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right) \tag{4}$$

where, w_{ij} is the connection weight between the i-th input and the j-th output, b_j is the bias term, and σ is the activation function (commonly used such as ReLU or Sigmoid function). Integrating the features obtained by the convolutional layer and pooling layer helps the fully connected layer to provide the final prediction result.

CNN's training is often accomplished using a backpropagation technique. Usually utilising a cross entropy loss function, the network computes the output by forward propagation, compares it with the real labels, computes the loss, and subsequently changes all weights and biases in the network via backpropagation algorithm. Each parameter's gradient is computed in backpropagation using the chain rule; the update rule is:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L \tag{5}$$

where, θ represents the parameters in the network (including weights and biases), η is the learning rate, L is the loss function, and $\nabla_\theta L$ is the gradient of the loss function on the parameters.

Widely applied in many disciplines including computer vision and audio processing, convolutional neural networks efficiently extract and integrate characteristics from input data by convolution operations, activation functions, pooling operations, and fully connected layers. Through local receptive fields and weight sharing mechanisms, the convolution operation lowers the number of parameters and improves computational efficiency; pooling techniques lower dimensionality and hence increase the resilience of the model. CNN has great learning capacity by automatically learning feature representations fit for various tasks using backpropagation technique.

2.2. **Long short term memory network.** Designed to address the issues of gradient vanishing and exploding experienced by conventional RNNs processing long sequence data, LSTM is a particular recurrent neural network (RNN) architecture [17]. As Figure

2 shows, LSTM can efficiently capture long-term relationships in time series by means of its special gating mechanism. Natural language processing, speech recognition, machine translation, and video analysis are just a few of the applications for LSTM that abound.

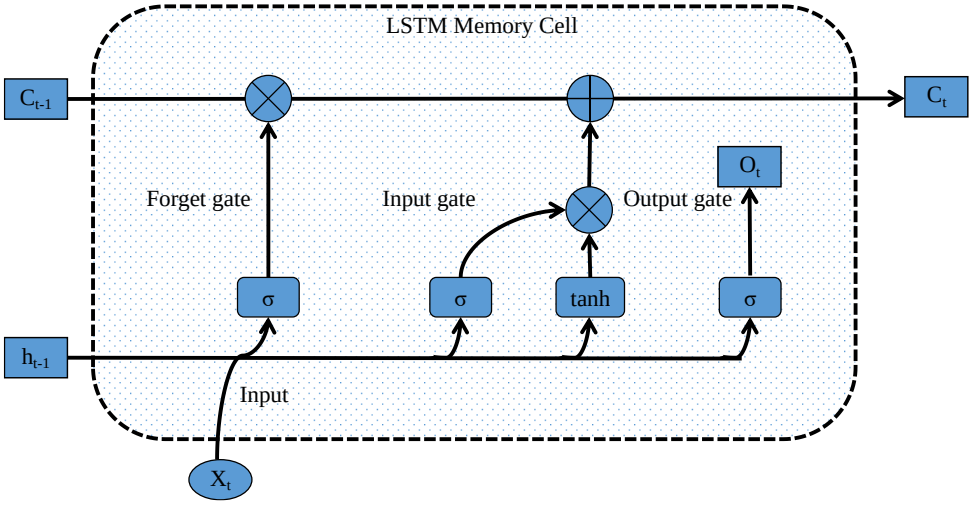

Figure 2. The architecture of LSTM

LSTM's fundamental architecture is to use gating techniques to regulate information flow, therefore preventing the gradient vanishing issue in conventional RNNs. Input gate, forget gate, and output gate are three primary gating methods LSTM presents. Furthermore, LSTM adds cell state as a "highway" for information flow, enabling long-distance network propagation of knowledge [18].

LSTM has the benefit in that it can efficiently record long-term dependencies in time series. LSTM may dynamically change the information flow by including forget gates, input gates, and output gates, therefore avoiding the gradient vanishing issue that conventional RNNs run across in lengthy sequence data. Furthermore, LSTM offers an information flow channel through cell state design, which helps to retain and distribute knowledge over an extended length of time thereby improving the memory capacity of the model.

LSTM exhibits the following traits over conventional RNNs: Long-term dependencies can be sufficiently captured in sequence data by LSTM via its gating mechanism. LSTM by use of information flow efficiently avoids the gradient vanishing and exploding issues in conventional RNNs. LSTM's gating structure causes the model's decision-making process to have a certain interpretability, which helps one to naturally grasp how the model uses data at every point.

Proposed to solve the constraints of conventional RNNs in handling long time series data, LSTM is a unique network architecture. LSTM can dynamically select to keep or discard information by including forget gates, input gates, and output gates, therefore capturing long-term dependencies. Widely used in many sequence modelling applications, LSTM shows considerable benefits notably in long-term sequence dependent challenges.

2.3. **Collaborative filtering recommendation.** Currently one of the most often used recommendation algorithms, Collaborative Filtering (CF) advises depending on user past behaviour (such as purchase records, reviews, clicks, etc.) or similarity between items [19]. Collaborative filtering's core tenet is "birds of a feather flock together, people flock together," that is, by use of group identification akin to users or objects to forecast users' interest in non-contact items. Working method of collaborative filtering suggests that two types of collaborative filtering recommendation systems exist: Item Based Collaborative

Filtering (IBCF) and User Based Collaborative Filtering (UBCF). Furthermore, model based and memory based collaborative filtering might be further divisions of collaborative filtering.

Collaborative filtering is fundamentally focused on recommending items of interest to consumers depending on their past behaviour or properties of the items. The fundamental presumption of the collaborative filtering system is that, should individuals have comparable ratings for item A and item B, their ratings for other products might likewise be similar going forward [20].

User based collaborative filtering techniques seek to identify a group of users most similar to the target user and project the target user's rating for non-contact items based on their ratings. The particular actions follow this: First, using their similarities, find like users. Typical approaches of similarity computation consist in cosine similarity, Pearson correlation coefficient, etc.

Often used to find the cosine value of the angle between two vectors, cosine similarity evaluates users or objects. For the rating vectors r_u and r_v of users u and v, cosine similarity is defined as:

$$sim(u,v) = \frac{\sum_{i \in I} r_u(i) \cdot r_v(i)}{\sqrt{\sum_{i \in I} r_u(i)^2} \cdot \sqrt{\sum_{i \in I} r_v(i)^2}} \tag{6}$$

where, $r_u(i)$ and $r_v(i)$ are the ratings of item i by users u and v, respectively, and I is the set of items rated by all users.

The Pearson correlation coefficient quantifies two variables' linear relationship. For the rating vectors r_u and r_v of users u and v, the Pearson correlation coefficient is defined as:

$$sim(u,v) = \frac{\sum_{i \in I} (r_u(i) - \bar{r}_u) \cdot (r_v(i) - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_u(i) - \bar{r}_u)^2} \cdot \sqrt{\sum_{i \in I} (r_v(i) - \bar{r}_v)^2}} \tag{7}$$

where, $\bar{r}_u$ and $\bar{r}_v$ are the average ratings of users u and v, respectively.

Following user similarity acquisition, one can forecast the rating of the target user for non-contact objects using the ratings of related users. The weighted average approach is the most often applied prediction equation:

$$\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{v \in N_u} sim(u,v) \cdot (r_v(i) - \bar{r}_v)}{\sum_{v \in N_u} |sim(u,v)|} \tag{8}$$

where, $\hat{r}_{u,i}$ is the predicted rating of user u for item i, $\bar{r}_u$ is the average rating of user u, N_u is the set of k users with the highest similarity to user u, $r_v(i)$ is the rating of item i by user v, and $\bar{r}_v$ is the average rating of user v.

For the target user u, based on a predicted rating of $\hat{r}_{u,i}$, the item with the highest predicted rating can be recommended. Recommendation algorithms generate recommendations for users by selecting users with high similarity to the target user and referencing their ratings of unrated items.

Though it emphasises the resemblance between objects rather than between persons, item based collaborative filtering is comparable to user based collaborative filtering. The fundamental phases of the item based collaborative filtering method consist as follows: Like user based collaborative filtering, preliminary calculations of item similarity must be done. Usually, item similarity is expressed as cosine similarity. Though the prediction is based on item similarity, the prediction scoring system resembles user driven collaborative filtering.

Calculating the similarity between individuals or items, the collaborative filtering recommendation system generates recommendations depending on past behaviour of like

persons or objects. Whereas item based collaborative filtering concentrates on the similarity between objects, user based collaborative filtering emphasises on the similarity between individuals. While collaborative filtering systems show outstanding suggestion performance, handling sparse data and cold start issues still presents significant difficulties.

3. **Personalized learning path recommendation method.** The core objective of this study is to combine multimodal learning, deep learning, and data analysis techniques to provide personalized learning path recommendations for children. Specifically, the system will integrate multimodal data from visual, auditory, and motion perception, and extract multidimensional features through deep learning models. Then, combined with time series analysis, the learning behavior patterns of children are identified, and finally personalized learning paths are recommended for them through collaborative filtering algorithms. The model framework of this method is shown in Figure 3.

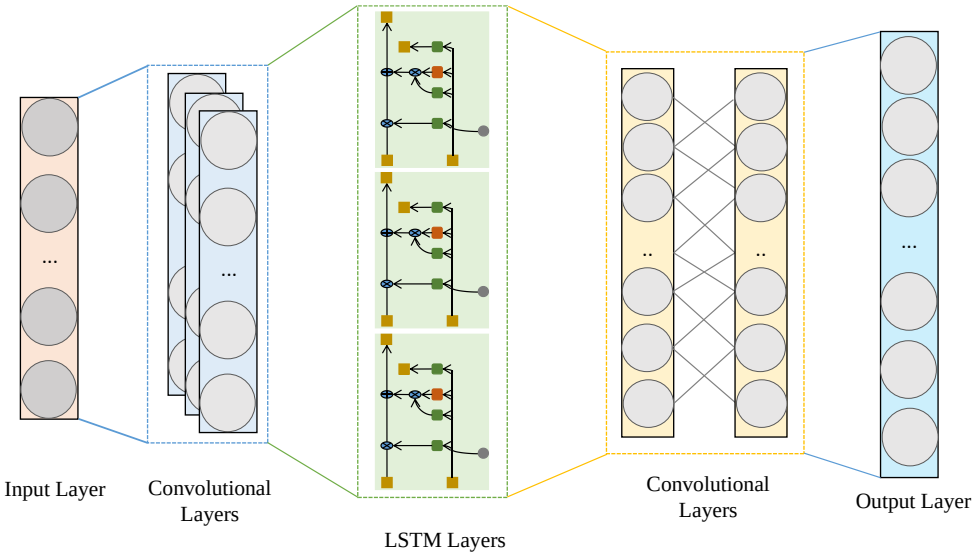

Figure 3. Method framework diagram

3.1. **Multi modal data feature extraction.** Multimodal learning involves integrating data from different senses to comprehensively analyze children's learning situations from multiple perspectives. For each sensory data (such as visual, auditory, and motion perception), use CNN for feature extraction.

For visual data, we first extract features from images using CNN. For example, changes in a child's facial expressions and posture can reflect their emotional state and help evaluate their learning progress. Assuming the input visual data is X_{visual}, after processing by the convolutional layer, the output is visual feature F_{visual}.

Convolution operation has a computation equation:

$$F_{visual} = ReLU(W * X_{visual} + b) \tag{9}$$

where, W is the convolution kernel, $*$ represents the convolution operation, X_{visual} is the input image, b is the bias term, and $ReLU$ is the activation function.

Auditory data usually exists in the form of audio signals, and the system collects audio data through a microphone. The audio signal is converted into a spectrogram through STFT, and then audio features are extracted through CNN. Assuming the audio data is X_{audio}, feature F_{audio} is obtained through convolutional layer processing:

$$F_{audio} = ReLU(W * X_{audio} + b) \tag{10}$$

where, X_{audio} is the input audio data, W is the convolution kernel, b is the bias term, and $ReLU$ is the activation function.

Action perception data is collected through sensors, recording children's movements and postures. Similar to visual and auditory data, motion perception data also requires feature extraction through convolutional neural networks. Assuming the action data is X_{motion}, feature F_{motion} is obtained through convolutional layers:

$$F_{motion} = ReLU(W * X_{motion} + b) \tag{11}$$

3.2. **Multimodal data fusion.** To obtain more comprehensive information on children's learning, we integrate data features from different senses. Assuming we have extracted features F_{visual}, F_{audio}, and F_{motion} from visual, auditory, and motion perception, we weight and fuse them into a global feature vector F_{multi}, with the specific formula as follows:

$$F_{multi} = \lambda_1 \cdot F_{visual} + \lambda_2 \cdot F_{audio} + \lambda_3 \cdot F_{motion} \tag{12}$$

where, λ_1, λ_2, and λ_3 are the weight coefficients of each modality, representing the importance of visual, auditory, and action data.

3.3. **Time series analysis.** Children's learning process often exhibits time series properties whereby emotions and learning progress change with time. We apply LSTM for time series data modelling for this aim. Long-term dependencies can be sufficiently captured by LSTM, which also helps to prevent the gradient vanishing issue in conventional RNNs.

Four main elements define every second of LSTM: input gate, forget gate, output gate, and cell state. These doors regulate the information flow within the network.

The function of the forget gate is to determine which information will be discarded from the cellular state. Its output is calculated by the sigmoid function using the current input x_t and the hidden state h_{t-1} from the previous moment, and the equation is as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{13}$$

where, W_f and U_f are the weight matrices of the forget gate, b_f is the bias term, σ is the sigmoid activation function, and the output value f_t ranges between $[0, 1]$, determining the degree of forgetting.

The input gate serves to decide which fresh information will be entered into the cell state. First, by means of the sigmoid function, the input gate decides which information will be updated at the present. Next, use the tanh function to generate a candidate value vector $\tilde{C}_t$, representing the candidate cell state at the current time. The equation is as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{14}$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{15}$$

where, W_i, U_i, and b_i are the weights and biases of the input gates, W_c, U_c, and b_c are the weights and biases for generating candidate cell states, and $tanh$ is the activation function.

Cell state C_t is updated at every moment based on the outputs of the forget gate and input gate. Specifically, the previous cell state C_{t-1} will be forgotten based on the output of the forget gate, while the input gate determines which new information will be written. The update equation for cell state is:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{16}$$

where, $*$ represents element-wise multiplication.

The output gate determines the hidden state h_t at the current time, which is the output of the network. The calculation of the output gate depends on the current cell state C_t and the output of the input gate i_t. Firstly, the sigmoid activation function determines

which parts of the cell state will be output, and then the tanh activation function is used to adjust the range of cell states, ultimately generating hidden state h_t. The equation is as follows:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{17}$$

$$h_t = o_t * \tanh(C_t) \tag{18}$$

where, W_o, U_o, and b_o are the weights and biases of the output gates. LSTM effectively captures and updates temporal information through these gating structures.

3.4. **Learning behavior pattern recognition and clustering.** We separate children's learning behaviour characteristics into several groups using the K-means clustering algorithm so as to detect their patterns. K-means clustering seeks to reduce the sum of squared distances between the sample and the cluster centre. Specifically, the equation is:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} 1_{\{y_i=k\}} \|x_i - \mu_k\|^2 \tag{19}$$

where, n is the number of samples, k is the number of clusters, x_i is the i-th sample, μ_k is the center of the k-th cluster, $1_{\{y_i=k\}}$ is the indicator function, and $1_{\{y_i=k\}}$ indicates that sample i belongs to cluster k.

By use of cluster analysis, we may detect children's behavioural patterns during the learning process (including positive learning, emotional fluctuations, etc.), therefore provide a foundation for tailored recommendations.

3.5. **Personalized learning path recommendation.** In this work, we implemented a user based collaborative filtering method for path recommendation for customised learning. Through analysis of commonalities, collaborative filtering forecasts children's interest in non-contact learning activities.

The core equation for collaborative filtering is:

$$\hat{r}_{u,i} = \frac{\sum_{j \in N_i} sim(i,j) \cdot r_u(j)}{\sum_{j \in N_i} |sim(i,j)|} \tag{20}$$

where, $\hat{r}_{u,i}$ is the predicted rating of user u for item i. N_i is the set of k items with the highest similarity to item i. Through this equation, we can predict each child's interest in the tasks they have not learned, and recommend personalized learning paths for them.

This approach uses CNN to extract features from many modalities together with multimodal learning and data analysis. For time series analysis, it uses LSTM; for tailored recommendations, it makes use of collaborative filtering systems. Combining these approaches helps one to precisely spot children's emotional changes and learning behaviour patterns, suggest customised learning paths for them, maximise learning results, and improve learning environments.

4. Experiment.

4.1. **Data set.** This work chose "EduKids Dataset" as the primary dataset for experimental purposes. Including multimodal data—visual (pictures of children's facial expressions), auditory (audio recordings of children's speech), and behavioural data—children's movements recorded using accelerometers and gyroscope sensors—this dataset is intended especially for early childhood education. 3000 samples make up this dataset, each sample with an emotional state (such as happy, uncertainty, anxiety, etc.) and learning progress (such as task completion time, accuracy, etc.) labelled. Data preparation consists in the following actions to guarantee the comparability and validity of data: Images for visual data are standardised and universally changed to 224×224 pixels. The behavioural data

is filtered and normalised, then transformed into a time series fit for the LSTM model; the audio signal first undergoes Short Time Fourier Transform (STFT) and subsequently becomes a spectrogram.

4.2. **Evaluation.** In order to comprehensively evaluate the performance of the proposed model, we have selected the following evaluation metrics:

One of the most often used classification evaluation metrics, accuracy gauges a model's classification correctness. The equation is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

where, TP stands for the actual number of cases; TN for the actual negative number; FP for the false positive number; FN for the false negative number.

F1 Score is the harmonic mean of precision and recall, particularly suitable for situations where data is imbalanced. The equation is:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{22}$$

where, Precision and Recall are:

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

Mean Square Error (MSE) is used to evaluate the prediction error of models in regression tasks, especially in learning progress prediction tasks. The equation is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{25}$$

where, y_i represents the true value, $\hat{y}_i$ represents the predicted value, and n is the number of samples.

4.3. **Experimental results and analysis.** We have chosen the following evaluation criteria to holistically assess the performance of the proposed model:

We evaluated the suggested model with the following baseline models to confirm its efficiency:

(1) Single modal CNN model: using only visual data (facial expression images) for emotion recognition and learning progress prediction.
(2) Single mode LSTM model: learns behavior analysis using only behavioral data.
(3) Multi modal learning DNN model: Combining visual, auditory, and behavioral data, using DNN for emotion recognition and learning progress prediction.
(4) Multi modal CNN+LSTM+collaborative filtering model: Combining all modal data, using collaborative filtering algorithms for personalized learning path recommendation.

This experiment evaluates the performance of four models in tasks such as emotion recognition, learning progress prediction, and personalized recommendation path generation by comparing their performance. The specific results are shown in Table 1 and Figure 4:

From the above experimental findings, one can observe that the model in this article performs well in all evaluation criteria, particularly in accuracy F1-score. Regarding coverage and recommendation accuracy, it stands far better than other baseline models. Low accuracy and F1 score of single-mode CNN model and single-mode LSTM model

Table 1. Model comparison results

Method	Accuracy	F1-Score	MSE
Single modal CNN model	0.774	0.732	0.112
Single mode LSTM model	0.742	0.717	0.134
DNN model for multimodal learning	0.838	0.809	0.090
Multimodal CNN+LSTM model	0.901	0.873	0.067

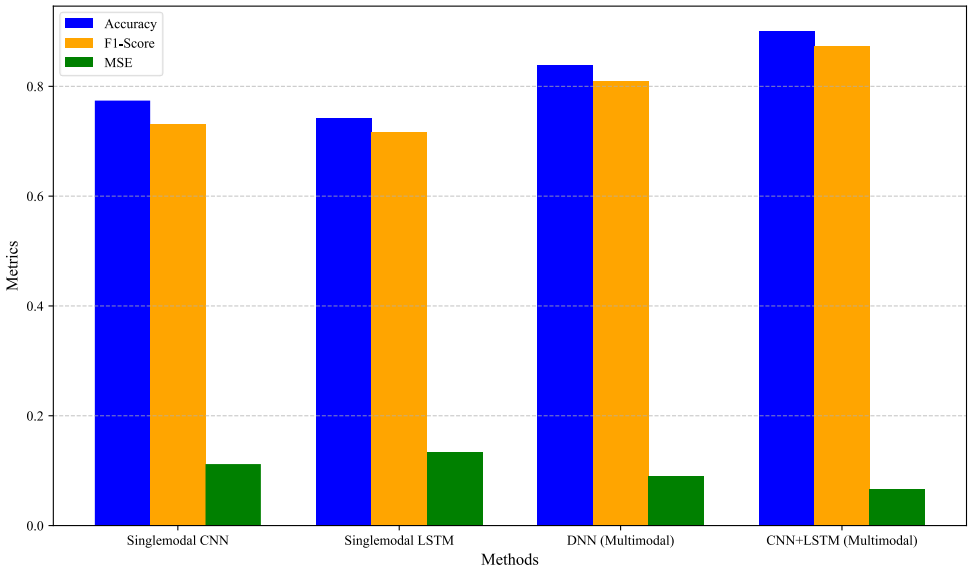

Figure 4. Model comparison results

show that the information of a single mode is insufficient to precisely recognise emotions and forecast learning development. Particularly for the single-mode LSTM model, its poor performance results from the absence of visual and auditory input even if it can manage the time series aspects of behavioural data. Combining visual and behavioural data, the multimodal DNN model works well to obtain high accuracy and F1 Score in emotion identification and learning progress prediction chores. This outcome suggests that children's richer learning elements can come from the combination of visual and behavioural data, therefore enhancing the model's performance. With an F1 Score of 0.873—the greatest performance among all models—the accuracy of the model in this article. Furthermore surpassing other models in terms of suggestion accuracy and coverage in individualised recommendation jobs is this model. This suggests that the implementation of collaborative filtering algorithms can offer customised learning paths depending on children's learning development and interests, therefore enhancing the accuracy and variety of recommendation systems.

The variation of the loss function with the number of iterations is shown in Figure 5. According to Figure 5, it can be seen that our method converges the fastest.

5. **Conclusion.** Aiming to maximise the learning experience and personalised learning path design of preschool children by integrating visual, auditory, and behavioural data and using deep learning technology, this paper suggests a personalised learning path recommendation method based on multimodal learning and data analysis for preschool

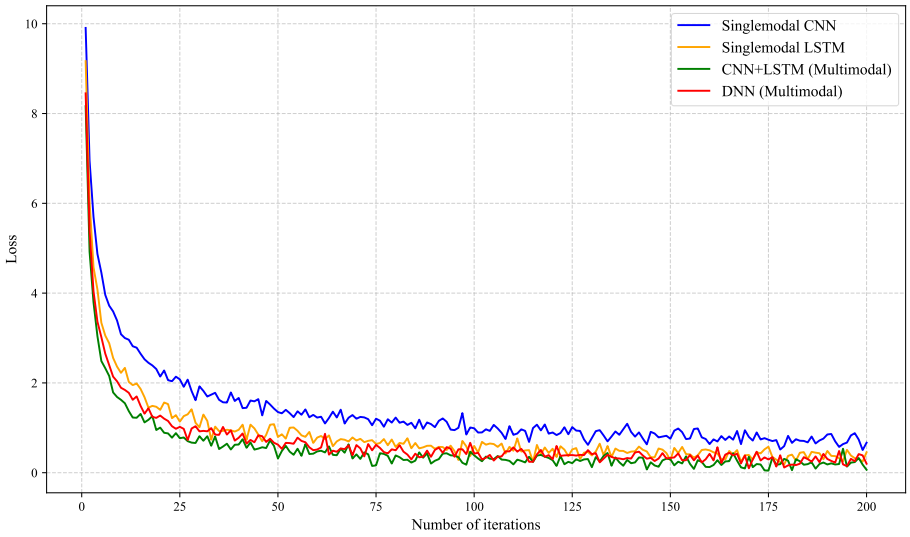

Figure 5. The variation of loss function with iteration times

Education. Through cooperative filtering algorithms, the model can fully explore multi-dimensional learning features of children and generate personalised recommendations that fit their emotional state and learning progress by combining multimodal CNN and LSTM. This work contributes by suggesting a new personalised learning path recommendation framework for preschool education and so enhancing the accuracy and diversity of personalised recommendations by multimodal learning techniques, so supporting the application of deep learning and recommendation systems in the field of education. Simultaneously, this study offers fresh concepts for the future development of intelligent teaching systems in preschool education, particularly in the domains of emotional recognition, learning progress tracking, and personalised learning recommendations—which have general relevance. Future work can investigate how to maximise the computational efficiency of the model to meet the real-time needs of big-scale online education platforms; in addition, how to integrate more diverse perceptual data (such as biological signals, emotional analysis, etc.) with existing data to improve the accuracy of recommendation systems is also a direction worthy of further research.

REFERENCES

[1] L. Mihelač, "Recommendation Systems, Parents, and Preschool Children: The Story Behind Digital Technology," *Revija-za Elementarno Izobrazevanje*, vol. 17, no. 2, pp. 155-170, 2024.

[2] S. Mu, M. Cui, and X. Huang, "Multimodal data fusion in learning analytics: A systematic review," *Sensors*, vol. 20, no. 23, 6856, 2020.

[3] S. V. Taylor and C. B. Leung, "Multimodal literacy and social interaction: Young children's literacy learning," *Early Childhood Education Journal*, vol. 48, no. 1, pp. 1-10, 2020.

[4] N. Kucirkova and R. Flewitt, "The future-gazing potential of digital personalization in young children's reading: views from education professionals and app designers," *Early Child Development and Care*, vol. 190, no. 2, pp. 135-149, 2020.

[5] M. Kim and S. O. Park, "Group affinity based social trust model for an intelligent movie recommender system," *Multimedia Tools and Applications*, vol. 64, pp. 505-516, 2013.

[6] R. Zhang, "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, 2022.

[7] A. Tripathi, T. Ashwin, and R. M. R. Guddeti, "EmoWare: A context-aware framework for personalized video recommendation using affective video sequences," *IEEE Access*, vol. 7, pp. 51185-51200, 2019.

[8] F. Ullah et al., "Deep edu: a deep neural collaborative filtering for educational services recommendation," *IEEE Access*, vol. 8, pp. 110915-110928, 2020.
[9] S. Bhaskaran and B. Santhi, "An efficient personalized trust based hybrid recommendation (tbhr) strategy for e-learning system in cloud computing," *Cluster Computing*, vol. 22, pp. 1137-1149, 2019.
[10] L. Crescenzi-Lanna, "Multimodal Learning Analytics research with young children: A systematic review," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1485-1504, 2020.
[11] Y. Zhou, C. Huang, Q. Hu, J. Zhu, and Y. Tang, "Personalized learning full-path recommendation model based on LSTM neural networks," *Information Sciences*, vol. 444, pp. 135-152, 2018.
[12] S. Xu and X. Yin, "Recommendation System for Privacy-Preserving Education Technologies," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, 3502992, 2022.
[13] H. J. Bang, L. Li, and K. Flynn, "Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning," *Early Childhood Education Journal*, vol. 51, no. 4, pp. 717-732, 2023.
[14] Y. Zheng, D. Wang, J. Zhang, Y. Li, Y. Xu, Y. Zhao, and Y. Zheng, "A unified framework for personalized learning pathway recommendation in e-learning contexts," *Education and Information Technologies*, vol. 143, pp. 1-38, 2024.
[15] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24-49, 2021.
[16] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit, "A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, 5930, 2023.
[17] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, "A survey on anomaly detection for technical systems using LSTM networks," *Computers in Industry*, vol. 131, 103498, 2021.
[18] H. N. Bhandari, B. Rimal, N. R. Pokhrel, R. Rimal, K. R. Dahal, and R. K. Khatri, "Predicting stock market index using LSTM," *Machine Learning with Applications*, vol. 9, p. 100320, 2022.
[19] H. Papadakis, A. Papagrigoriou, C. Panagiotakis, E. Kosmas, and P. Fragopoulou, "Collaborative filtering recommender systems taxonomy," *Knowledge and Information Systems*, vol. 64, no. 1, pp. 35-74, 2022.
[20] F. Fkih, "Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 7645-7669, 2022.