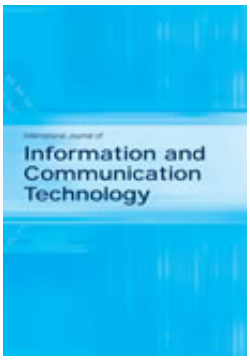


International Journal of Information and Communication Technology

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

A multimodal health data fusion and deep analysis approach in smart campus systems

Fang Fu, Zhiguang Li

DOI: 10.1504/IJICT.2025.10071390

Article History:

Received:	15 April 2025
Last revised:	28 April 2025
Accepted:	29 April 2025
Published online:	11 June 2025

A multimodal health data fusion and deep analysis approach in smart campus systems

Fang Fu*

School of Sport and Physical Education,
North University of China,
Taiyuan 030051, China
Email: lionlian007@163.com
*Corresponding author

Zhiguang Li

School of Advanced Imaging Science, Multimedia and Film,
Chung-Ang University,
Seoul 06974, South Korea
Email: makefilmcau@126.com

Abstract: Health monitoring, a key component of the smart campus system, involves multimodal health data. Aiming at the problem of insufficient intermodal interactivity in existing research, we first vectorise the multimodal health data, such as audio, image and text, and design a multiscale convolutional neural network (MCNN) to extract the multimodal data features and carry out statistical pooling to obtain the standard deviation, maximum value and average value of the feature vectors. Then, the dense attention mechanism (DAM) is designed to realise the interactive fusion of multimodal data, the multivariate Gaussian distribution is utilised to classify the health states, and the multilayer perceptron is combined to construct the health data analysis algorithm. The experimental results show that the fusion efficiency of the proposed method is greater than 85% and the classification accuracy reaches 95.07%, which significantly improves the monitoring of multimodal health data.

Keywords: smart campus; health monitoring; multimodal data fusion; multiscale convolutional neural network; MCNN; dense attention mechanism; DAM.

Reference to this paper should be made as follows: Fu, F. and Li, Z. (2025) 'A multimodal health data fusion and deep analysis approach in smart campus systems', *Int. J. Information and Communication Technology*, Vol. 26, No. 17, pp.147–162.

Biographical notes: Fang Fu received her Master's degree from the Shanxi University in June 2008. She is currently working in the School of Sport and Physical Education at North University of China. Her research interests include physical education and sports management.

Zhiguang Li received his Bachelor's degree from the Zhengzhou University in June 2020. He is currently studying for a PhD degree at the Graduate School of Chung-Ang University, Korea. His research interests include machine learning and artificial general intelligent control.

1 Introduction

In the context of the deep integration of education informatisation and artificial intelligence technology, the smart campus system has become an important carrier to enhance the efficiency of education management and optimise the allocation of resources (Dong et al., 2020). As one of the core functional modules of the smart campus, the student health monitoring system continuously collects multimodal health information of students through multi-source terminals. However, these data are often characterised by strong heterogeneity, complex dimensions, and fuzzy dynamic associations, making it difficult for traditional single-modal analysis methods to comprehensively reveal the multidimensional association patterns of students' health status and to achieve accurate early warning of personalised health risks (Wang et al., 2024). Recently, multimodal health data fusion and in-depth analysis methods have emerged, which are capable of integrating different types of health data and mining the hidden information behind the data so as to provide powerful support for campus health management decisions (Xie et al., 2024). Through this innovative approach, it is expected to achieve early warning of students' health risks, personalised health interventions, and optimisation of the campus health environment, thus promoting the smart campus to play a greater role in safeguarding students' physical and mental health (Anagnostopoulos et al., 2021).

Roda-Sanchez et al. (2023) designed a set of integrated campus design solutions oriented to the innovative service ecosystem of smart campuses, which provides new ideas for the construction of smart cities driven by big data and digital twins. Li (2021) investigates the objectives, technical framework, applications, and application effects of a smart campus health and wellness management system. Huang et al. (2024) used 360° video technology to collect students' multimodal health data, such as physiological data, behavioural data, psychological data, and environmental data, to support accurate campus management. John et al. (2021) used wearable and implantable devices to automatically capture, encode, and process multimodal student health data to characterise students' moods and expressions and to predict certain instructional activities of teachers. Traditional campus management faces a number of challenges, including fragmented data, information silos, and insufficient basis for decision-making, which constrain the improvement of campus management.

Machine learning can automatically extract features from data, reducing the reliance on domain knowledge and providing new ideas for smart campus management. Cai (2023) used BP neural networks to identify and extract behavioural characteristics of teachers and students in classroom teaching and learning based on smart classroom data, and adopted an artificial intelligence engine to automatically label classroom teaching behaviours. Haleem et al. (2023) used deep learning algorithms to analyse students' behaviour, physiology, psychology and other health data to comprehensively reflect students' learning and teachers' teaching behaviours, and help teachers optimise the teaching process. Liang et al. (2021) combined students' audio and text modal data, learned audio features by principal component analysis (PCA) and support vector machine (SVM), and extracted key words in the text for students' health monitoring, and achieved good monitoring results. Jafari et al. (2018) used convolutional neural networks (CNNs) to process text and video modalities separately, and then combined the results of these analyses through a logistic regression model to improve the fusion efficiency. Wang (2024) integrated speech and video modalities and utilised a bidirectional long and short-term memory network (Bi-LSTM) and SVM to improve the prediction accuracy.

Song et al. (2023) used an attentional mechanism to splice different modal data two by two, and then passed this spliced information to Bi-LSTM to learn cross-modal contextual correlations. Chao et al. (2024) established a multimodal student health monitoring model with composite hierarchical fusion by combining temporal convolutional networks (TCNs) and soft attention mechanisms to improve prediction accuracy. Fang et al. (2023) used a dynamic augmentation approach to extend the difference between textual modalities and other modalities and reduce the redundancy of other modalities, capturing the context of multimodal data through a bidirectional attention mechanism, but failing to capture multilevel interactions between modalities.

Through a comprehensive analysis of multimodal data fusion and classification monitoring methods in the smart campus system, it can be seen that the existing research exists the problems of insufficient inter-modal interaction and in the feature expression ability, in order to solve these problems, this paper proposes a highly efficient multimodal health data fusion and in-depth analysis method in the smart campus system. The innovativeness of this research is reflected in the following four aspects.

1 Deep learning algorithms were introduced to pre-process multimodal health data. Skip-Gram word embedding method is utilised to obtain word vectors for text, the VGG16 model pre-trained by ImageNet is introduced to obtain image vectors, and Mel frequency cepstrum coefficients (MFCC) is utilised to obtain audio vectors.

2 Multiscale convolutional neural network (MCNN) is designed to extract features from multimodal health data. By stacking multiple dense attention mechanism (DAM) layers to capture and fuse different modal health data, each layer not only refines the key information at its own level, but also provides the semantic information for the next layer, which significantly improves the model's capability to perceive the nuances in the characteristics of the health data.

3 After feature fusion, students' health status was classified and analysed in depth using multivariate Gaussian distribution. Finally, based on the results of the health status analysis, schools, teachers and parents can take appropriate interventions according to the results of the students' health status analysis, so as to improve the efficiency of campus management.

4 A large number of comparative experiments have been conducted on real datasets, and the outcome indicates that the proposed method has high data fusion efficiency and classification accuracy, provides theoretical support and technical implementation paths for the smart campus health management sub-system, and has practical value for promoting students' physical and mental health management, disease prevention and control, and personalised intervention.

2 Relevant technologies

2.1 Multimodal data fusion methods

Multimodal data fusion refers to the integration and analysis of data from various modalities to make full use of the complementary information of various modal data to improve the accuracy and comprehensiveness of understanding and cognition (Nemati et al., 2019). According to the fusion level, it can be categorised into data-level

integration, characteristic-level integration and decision-level integration. Data in intelligent campus systems has different spatio-temporal resolutions and semantic levels. The attention mechanism automatically focuses on the modality most relevant to the current task by dynamically calculating the weights of each modality feature, avoiding the static limitations of traditional weighted fusion.

1 Data-level integration characterises the model with sufficient data information through correlation and processing of raw data. If the original data has a large error, it will lead to a large deviation in the decision-making result.

2 Feature-level fusion is based on the potential features of the original data, compared with data-level fusion; both reduce the difficulty of data fusion, fusion model real-time effective enhancement.

3 Decision-level fusion is oriented to the integration of decision-making results at the end of data processing, and compared with a single data source, the decision-level fusion results are more accurate, robust, and fault-tolerant.

Data-level fusion requires strict spatiotemporal synchronisation, (e.g., video frame and audio sampling alignment), and the heterogeneous nature of devices in campus scenarios, (e.g., different sampling rates of cameras and IoT sensors) can easily lead to fusion failure. Feature-level fusion may ignore high-order correlations between modalities. Decision-level fusion processes each modality independently and supports asynchronous data, making it suitable for distributed systems in campuses.

2.2 Convolutional neural network

CNN is a neural network architecture based on the multilayer perceptron (MLP) design, which uses convolutional operations to extract and learn spatial features in a multilayer network. Unlike traditional neural networks, the neurons in each layer of a CNN are organised in a three-dimensional structure, which is more suitable for processing data such as images. CNN includes convolutional level, activation level, pooling level and fully connected level (Kuo, 2016) as shown in Figure 1. CNNs are classic deep learning models that demonstrate unique advantages when processing data with local correlations and spatial structures. Compared with other mainstream models (such as GANs and RNNs), CNNs automatically extract local features by sliding convolution kernels across local regions of the input data, significantly reducing the number of parameters.

Taking 2D convolution as an example, assuming that the size of the input feature map is $W_{in} \times H_{in} \times D_{in}$ and the size of the output feature map is $W_{out} \times H_{out} \times D_{out}$, the convolution is calculated as follows.

$$\begin{cases} W_{out} = \left(W_{in} + 2p - w\right)/s + 1 \\ H_{out} = \left(H_{in} + 2p - h\right)/s + 1 \\ D_{out} = k \end{cases} \tag{1}$$

where $w \times h$ is the width and height of the convolution kernel, k is the total number of convolution kernels, s is the step size, and p is the padding, which is used to control the spatial dimension of the output feature map.

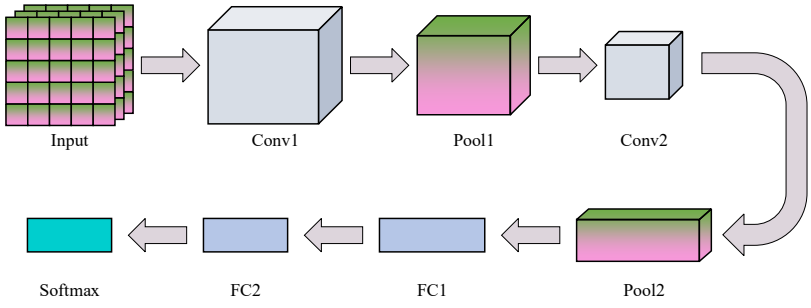
The pooling layer effectively reduces the spatial dimensionality of the characteristic picture, while reducing the amount of parameters and computational burden that the

model needs to handle (Li et al., 2021), and the maximum pooling is calculated as follows, where f is the width and height of the nucleus.

$$\begin{cases} W_{out} = (W_{in} - f)/s + 1 \\ H_{out} = (H_{in} - f)/s + 1 \end{cases} \tag{2}$$

The fully connected level maps the captured high-level characteristics to the output, performing a linear transformation $y = Wx + b$, where y is the output and W and b are the weights and bias, respectively.

Figure 1 The model structure of CNN (see online version for colours)



2.3 Attention mechanism

The attention mechanism is a technique used to enhance the attention of a neural network model to the input data by assigning a weight to each input location, focusing the model's attention on the part that is relevant to the task at hand. In this way, the model can better utilise the data in the input sequence and adaptively adjust the allocation of attention in different contexts.

In the attention mechanism, Q, K, and V (query-key-value) are used to compute the attention weights (Lu et al., 2023) for mapping the query vector (query), key vector (key), and value vector (value) to the attention scores, which are computed as follows.

$$Attention(Q, K, V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{3}$$

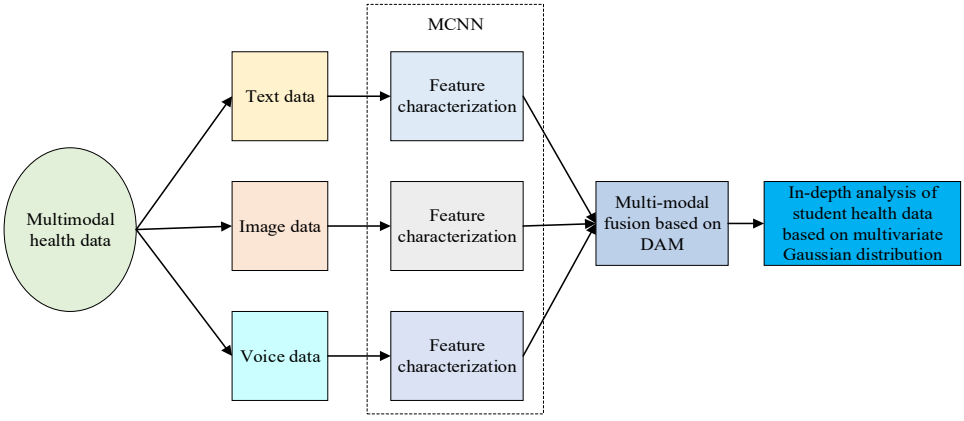
where d_k is the dimension of the key vector. This procedure computes the attention weights between multiple query vectors and key vectors at the same time and generates the corresponding context vectors or weighted sum representations.

3 General framework of health data analysis methods based on multimodal data

The smart campus system collects and analyses student health data through a variety of sensors, wearable devices, and information systems in order to provide personalised health management and intervention. These data are often characterised by rapid growth and multiple modalities, and it is difficult to ensure the accuracy of all modalities using

statistical data analysis algorithms. Therefore, it is necessary to design different data characterisation models for different modalities of health data while ensuring the accuracy. The framework of the proposed health data analysis algorithm is shown in Figure 2.

Figure 2 The framework of the proposed health data analysis algorithm (see online version for colours)



Due to the increase in the number of testing programs and the sophistication of instruments, student health data has expanded from a single digital and text type to a multimodal form with images, audio and text. However, the feature extraction methods for different modalities are very different, so it is necessary to classify the data into three modalities: text, image, and audio, and construct the feature characterisation model for each modality. Deep learning technology can be used to analyse the logical patterns embedded in data by simulating the way the human brain thinks, and unsupervised methods can avoid the subjective bias of manual annotation. In the paper, the data feature representation model of MCNN is constructed for three modal health data. The AM is also designed to interact and fuse the multimodal health data, and finally the health status of each type of student is analysed in depth using multivariate Gaussian distribution. Schools, teachers and parents can take appropriate interventions according to the results of students' health status analysis, which provides a reference for the development of personalised health management programs.

4 Multimodal health data fusion based on convolutional neural network and improved attention mechanism

4.1 Multimodal health data pre-processing

Intending to the issue that existing multimodal data fusion methods have insufficient intermodal interaction, which leads to inefficient fusion, a multimodal health data integration approach relied on MCNN and improved attention mechanism is suggested. Firstly, multimodal data such as audio, image and text are vectorised, MCNN is designed to extract the features of multimodal data, and finally, DAM is designed to establish

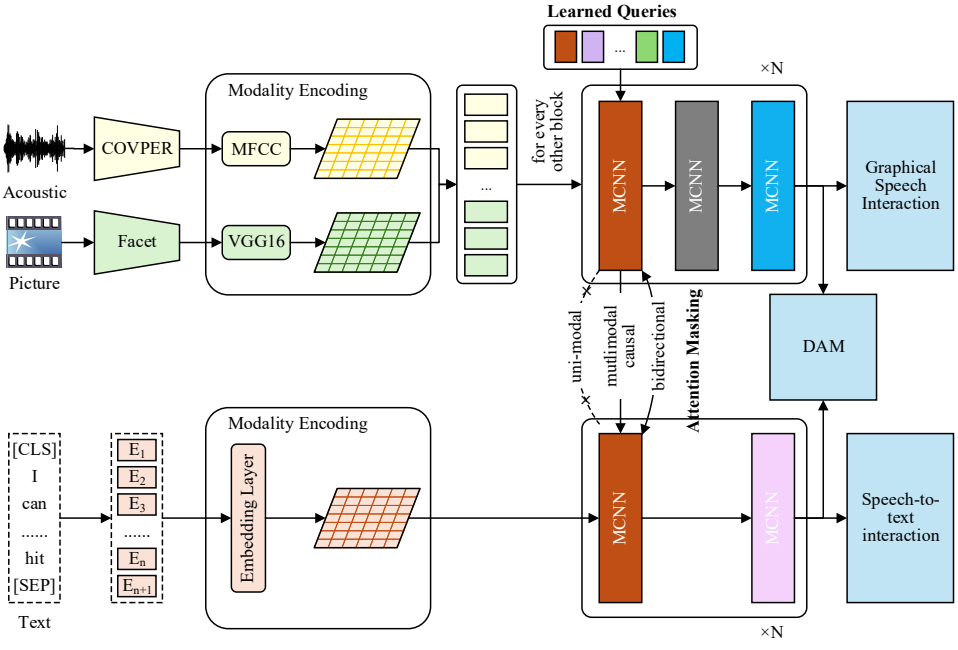
dense, multilevel interactions between modalities, which improves the intermodal complementarity and fusion efficiency, and realises the integration of deeper information.

The purpose of pre-processing multimodal health data is to vectorise the multimodal data. For textual health data, the jieba tool is used for word segmentation, and the Skip-Gram word embedding method is used for unsupervised learning of textual data and word headings to obtain the word vectors and their headings in the text. The model represents each word as a vector of words in the centre and background to calculate the conditional probability between the centre and background words.

$$p\left(w_o \mid w_c\right)=\frac{\exp\left(v_o^T v_c\right)}{\sum_{i=1}^{N}\exp\left(v_i^T v_c\right)} \tag{4}$$

where w_c is the centre word, w_o is the background word, v_c is the word vector of the centre word, v_o is the word vector of the background word, v_i is the word vector of the i^{th} background word in the dictionary, and N is the dictionary size.

Figure 3 Multimodal health data fusion process based on MCNN and DAM (see online version for colours)



For image-type health data, the ImageNet pre-trained VGG16 model (Ye et al., 2021) is introduced, and the last pooling layer (pool5) of this model is utilised for vectorisation of image data. The fixed-dimensional output of VGG16 eliminates differences in the size and resolution of the original images. The image vector output by pool5 in VGG16 is $7 \times 7 \times 512$, where 7×7 is the number of vectors and 512 is the vector dimension. Let $img = \{img_1, img_2, \ldots, img_n\}$ denote n images corresponding to the text, the image vector obtained by pool5 extraction is v_i. Input v_i into the fully connected level and nonlinear activation, the final image vector is obtained as follows.

$$v_{img} = \tanh\left(W^T v_i + b\right) \tag{5}$$

where tanh is the hyperbolic tangent activation function, W is the weight matrix, b is the bias vector, and d is the vector dimension.

For the audio type of health data, the discriminatory components of the audio data are extracted using MFCC. The audio is first converted to Mel frequency as follows.

$$mel(f) = 2\,595 \lg(1 + f / 700) \tag{6}$$

where f is the audio data. The audio vector is then obtained by performing Fourier transform, logarithmic operation and Fourier inverse transform on the audio data.

4.2 Multi-modal health data feature extraction based on multi-scale CNNs

After obtaining the multimodal health data vectors, this paper designs MCNN for multimodal data feature extraction. Let S be the text word vector, n be the text length, and k be the size of convolution kernel. For a certain convolution kernel w, the convolution operation is performed sequentially in the word vector matrix with k convolution kernels to obtain the text feature vectors one by one.

$$c_j = \text{Relu}\left(w^T * S_{j:j+k-1} + b\right) \tag{7}$$

where Relu is the activation function, * is the convolution operation, j is the number of features, and its value range is $[1, n - h + 1]$. In view of the above analysis, we can get the text feature vector is $V_t = \{c_1, c_2, \ldots, c_{n-k+1}\}$.

Similarly, for audio data and image data, the mean, maximum and standard deviation are also commonly used statistical features, while three types of pooling operations are used to portray the statistical features of multimodal data to obtain the statistical pooling vector $V^{SP} = \{V_t^{\mu}, V_e^{\mu}, V_s^{\mu}\}$, where μ is the pooling operation, which includes maximum pooling max, mean pooling avg and standard deviation pooling std; V_t, V_e and V_s are the text, image and audio feature vectors output from the feature extraction layer, respectively.

4.3 Multimodal health data fusion based on improved attention mechanism

Multimodal fusion is a key issue in student health data analysis, and this paper designs a DAM to create intensive, bi-directional interactions between modalities, as indicated in Figure 4. DAM captures and synthesises health data from different modalities by stacking multiple dense synergetic attention layers. Taking text and image as an example, this paper adopts the method in the literature (Al-Tameemi et al., 2023) to project V_t and V_e to multiple low-dimensional spaces respectively, and interact bimodally on multiple low-dimensional spaces to generate multiple attention graphs. In this way, interactions captured in different lower dimensions reinforce the correlation between signals, and the resulting multiple attention maps are averaged and fused, a process that allows the model to capture and exploit complex interactions between modes while maintaining dimensional control. The number of low-dimensional spaces is represented by h, and d_h is the number of dimensions of the low-dimensional space. The shared similarity matrix in the i^{th} low-dimensional space is shown as follows.

$$A_l^{(i)} = \left(W_{\tilde{V}_t}^{(i)} \tilde{V}_t\right) \cdot \left(W_{\tilde{V}_e}^{(i)} \tilde{V}_e\right) \tag{8}$$

where $W_{\tilde{V}_t}^{(i)}$ and $W_{\tilde{V}_e}^{(i)}$ are the linear weights of the text modality and image modality projected to the i^{th} low-dimensional space, respectively, and $A_l^{(i)}$ stores the interaction information of the two modalities.

Normalise the rows of the shared similarity matrix, to obtain the attention graph $A_{\tilde{V}_t}^{(i)}$ used to map the text on the image modal time nodes as shown in equation (9). The normalised columns then yield the text-to-image attention map as shown in equation (10). Next, by averaging multiple obtained attention maps, as shown in equation (11) and equation (12), a combined attention map is finally formed.

$$A_{\tilde{V}_t}^{(i)} = softmax\left(\frac{A_t^{(i)}}{\sqrt{d_h}}\right) \tag{9}$$

$$A_{\tilde{V}_e}^{(i)} = softmax\left(\frac{A_l^{(i)T}}{\sqrt{d_h}}\right) \tag{10}$$

$$A_{\tilde{V}_t} = \frac{1}{h}\sum_{i=1}^{h} A_{\tilde{V}_t}^{(i)} \tag{11}$$

$$A_{\tilde{V}_e} = \frac{1}{h}\sum_{i=1}^{h} A_{\tilde{V}_d}^{(i)} \tag{12}$$

Then the text and image modal fusion representations $\hat{V}_t = V_t A_{\tilde{V}_t}^T$ and $\hat{V}_e = V_e A_{\tilde{V}_e}^T$ are computed respectively, and finally, $\hat{V}_t$ and $\hat{V}_e$ are spliced with V_t and V_e of the previous layer, and the spliced features are projected into the d-dimensional space through a linear network with ReLU activation and residual concatenation. The final modal update process is shown in equation (13) and equation (14).

$$V_t^{l+1} = ReLU\left(W_{V_t}\begin{bmatrix} V_t \\ \hat{V}_t \end{bmatrix} + b_t\right) + V_t \tag{13}$$

$$V_e^{l+1} = ReLU\left(W_{V_e}\begin{bmatrix} V_e \\ \hat{V}_e \end{bmatrix} + b_e\right) + V_e \tag{14}$$

where W_{V_t} and W_{V_e} are weight coefficients and b_t and b_e are biases, respectively.

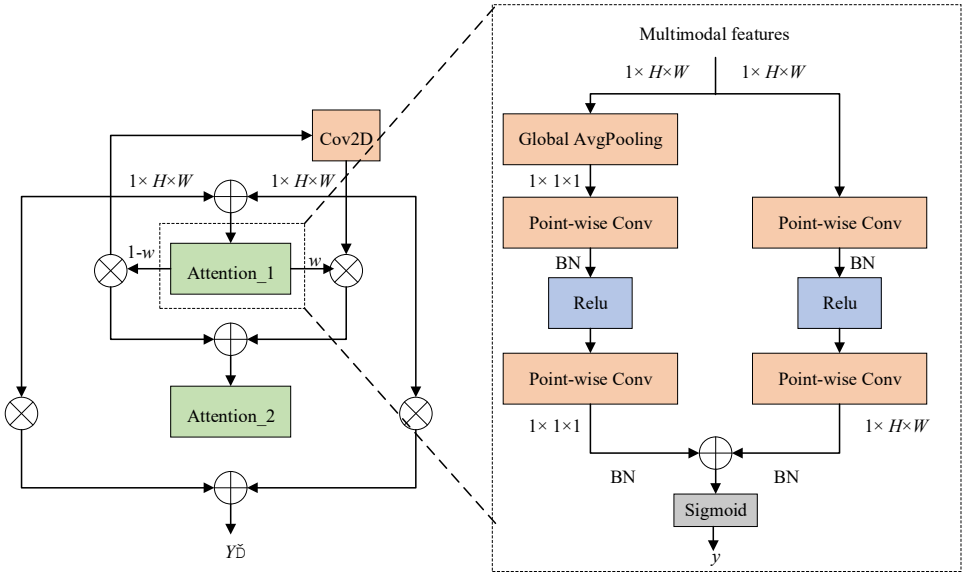
Through residual connection, information is allowed to be transmitted more directly in the network, which helps to reduce the problem of disappearing gradients, improve the effect of gradient propagation, and contribute to the stability of model training. Input the output T and V of the DAM layer into the MLP to gain the final fusion result, as shown in equation (15).

$$y_{te} = \sigma\left(MLP\left(\begin{bmatrix} V_t \\ T_e \end{bmatrix}\right)\right) \tag{15}$$

where y_{te} is the interaction result of the text-image modality and σ is the activation function. Similarly, the results of the modal interactions of text-to-speech and image will be obtained as y_{ts}, y_{es}. y_{te}, y_{ts} and y_{es} are weighted to obtain the fusion feature y_m, as shown in equation (16), where k_1, k_2 and k_3 are the weights of different modalities.

$$y_m = k_1 y_{te} + k_2 y_{ts} + k_3 y_{es} \tag{16}$$

Figure 4 The structure of DAM (see online version for colours)



5 Student health data analysis based on multivariate Gaussian distribution and deep learning

After obtaining the multimodal health data fusion feature results of students, this paper constructs a health data analysis model based on multivariate Gaussian distribution theory. The multivariate Gaussian distribution has become an important tool in multidimensional data analysis due to its excellent mathematical properties, concise parameters, efficient computation, and wide applicability. Firstly, the fused feature sequences of text, image and audio modalities are averaged for p.

$$p = \frac{1}{m}\sum_{i=1}^{m} y_i \tag{17}$$

The covariance matrix A of the features can be obtained from the mean value p as follows.

$$A = \frac{1}{m}\sum_{i=1}^{m}(y_i - p)(y_i - p)^T \tag{18}$$

The final probability value $p(x)$ of the multivariate Gaussian distribution can be obtained as follows.

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|A|^{\frac{1}{2}}}\exp\left((x-p)^T / A(x-p)\right) \tag{19}$$

Based on the above formula, a suitable Gaussian probability distribution function is selected based on the probability of occurrence of various diseases; then the fusion eigenvalues calculated above are used and inputted into the MLP to obtain a series of relevant parameters of the Gaussian probability distribution function, and the thresholds for the division of each student's health status are then obtained, and the final test samples are used for testing in order to verify whether the model has an effect of health data monitoring.

Based on the results of health data monitoring and analysis, schools can send relevant alert messages to school administrators, teachers and parents to remind them to pay attention to students' health conditions according to the Smart Campus Health Management Alert Module. Teachers can adjust their teaching plans and methods according to the health conditions of individual students, so as to reduce students' learning pressure. Parents can cooperate with the school to pay attention to the living habits and psychological state of the students and give them more care and attention.

6 Experimental results and analyses

This experiment was conducted on an NVIDIA Tesla P100 PCle 16 GB graphics card using the Pytorch framework, Pycharm environment, programming language Python 3.7 and CUDA version 11.0. The student health data in the smart campus system collected in the literature (Liang and Chen, 2018) is selected as the experimental dataset, which contains 21,678 multimodal health data of 4,289 students, such as physiological data, behavioural data, psychological data, and environmental data stored in the form of text, image, or audio. 60% of this dataset is used as the training set, 20% as the validation set, and 20% as the test set. The experiments are optimised using the Adam optimiser, and the batch size of the training and testing phases is 64, with a total of 300 rounds of training, and the original studying rate is set to 0.001.

As can be seen from Figure 5, the correlation coefficient among the predicted and actual values of the proposed OURS method on the training sample for the 100-th training is 0.99801, the correlation coefficient between the predicted and actual values on the training set is 0.981, and the correlation coefficient between the predicted and actual values on the whole training sample is 0.99801, and the correlation coefficient between the predicted and actual values on the testing set is 0.981.

In addition to analysing the monitoring results of the OURS method, this paper also compares and analyses the fusion efficiency of the OURS method with MIFM (Song

et al., 2023), TCN-SAM (Chao et al., 2024), and DEBAM (Fang et al., 2023), and the results are shown in Table 1. When the training time is 10 s, the fusion efficiency of OURS is 33%, 15%, and 4% higher than that of the MIFM, TCN-SAM, and DEBAM methods, respectively, and when the training time is 50 s, the fusion efficiency of OURS is 91%, which is 48%, 19%, and 5% higher than that of the MIFM, TCN-SAM, and DEBAM methods, respectively. MIFM uses a traditional attention mechanism for simple splicing fusion of multimodal data features, which assigns weights to all input features, including noisy or irrelevant features, resulting in inefficient fusion. Although TCN-SAM fuses multimodal features through a soft-attention mechanism, the soft-attention mechanism usually assigns non-zero weights to all input features and lacks sparsity, which may lead to difficulties in focusing the model on key features and reduce the effectiveness of feature fusion. DEBAM fuses multimodal features through a two-way interactive attention mechanism, but it needs to store the attention weight matrix between two feature sets, and the memory occupation is high. To summarise, the fusion efficiency is low using the benchmark method, while the fusion efficiency is high using the OURS method.

Moreover, accuracy, Macro-F1, mean absolute error (MAE), mean squared error (MSE), and AUC values were used in this paper to compare the monitoring performance of different methods, and the results are shown in Figure 6. The Accuracy and Macro-F1 of OURS are 0.9507 and 0.9379, respectively, which are 23% and 19.58% improved compared to MIFM, 14.17% and 10.72% improved compared to TCN-SAM, and 3.63% and 3.54% improved compared to DEBAM. Comparing the prediction error metrics again, the MAE and MSE of OURS are at least 28.07% and 29.49% lower compared to the other three methods, respectively. AUC is the area of the offline surface of the ROC, which takes values ranging from 0 to 1. It also takes into account the recall rate (TPR) and false positive rate (FPR) of the model, and is able to fully assess the monitoring performance of the model. The AUC values of OURS, MIFM, TCN-SAM, and DEBAM were 0.9811, 0.8736, 0.9258, and 0.9525, respectively, and OURS was more effective in monitoring health data. OURS not only considers three modalities of health data and designs MCNN for multi-scale feature extraction of multimodal health data, but also innovatively proposes DAM to realise the deep interaction and fusion of multimodal data, which significantly improves the monitoring effect of multimodal health data.

Table 1 Fusion efficiency of different methods

Fusion time/s	*MIFM*	*TCN-SAM*	*DEBAM*	*OURS*
10	56%	74%	85%	89%
20	62%	76%	84%	92%
30	59%	71%	81%	93%
40	41%	78%	82%	87%
50	43%	72%	86%	91%
60	52%	75%	80%	95%
70	51%	77%	83%	89%
80	50%	70%	81%	90%

Figure 5 Effectiveness of the fit between predicted and actual values (see online version for colours)

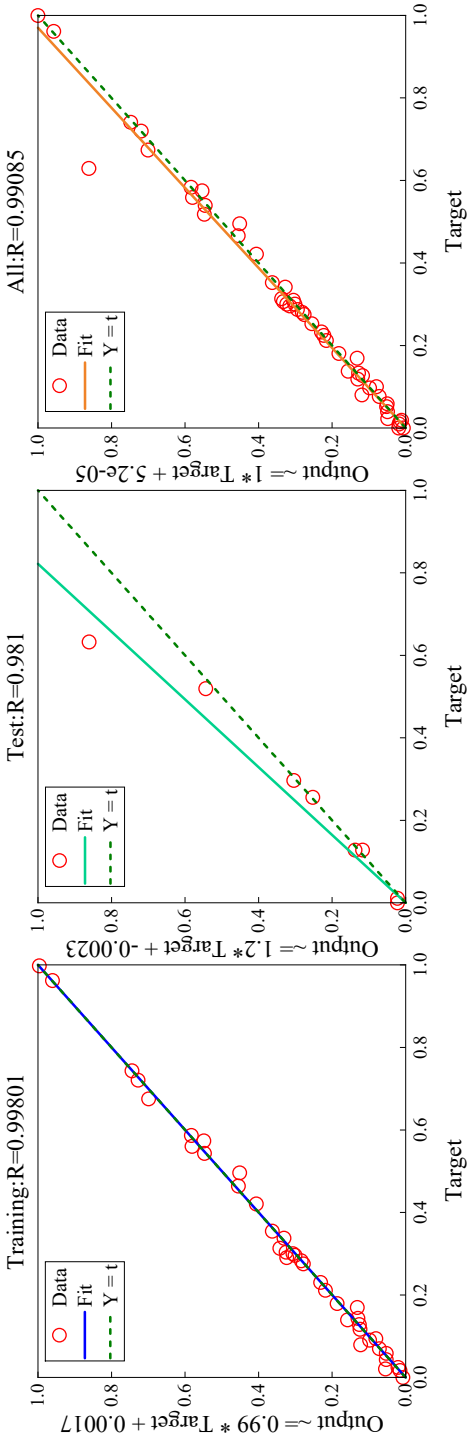
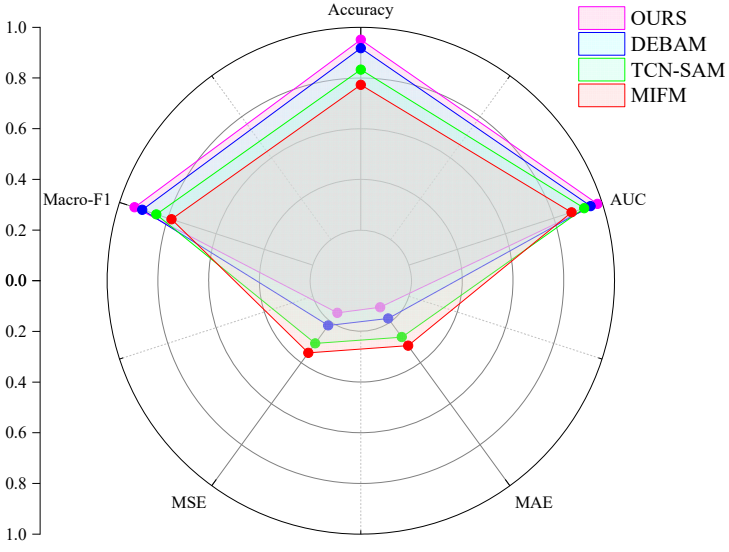


Figure 6 Performance indicators for monitoring and analysis of different methods (see online version for colours)



7 Conclusions

Intending to the current problem of insufficient interaction of multimodal data and weak feature expression ability in the smart campus system, which leads to the inefficiency of multimodal data fusion, this paper firstly obtains the word vectors of the text by using the Skip-Gram word embedding method, introduces the ImageNet pre-trained VGG16 model to obtain the image vectors, and obtains the audio vectors by utilising the Mayer frequency cepstrum coefficients (MFCC), and then CNN models with different convolutional kernels are designed for text, image and audio multimodal health data for feature extraction. The introduction of DAM establishes a dense, bi-directional interaction between modalities, and significantly improves the model's ability to perceive subtle differences in the characteristics of health data by stacking multiple dense synergetic attention layers to capture and fuse health data from different modalities. After feature fusion, students' health status was divided and analysed in depth using multivariate Gaussian distribution. Finally, based on the results of the analysis of health status, schools, teachers and parents can take appropriate interventions according to the results of the analysis of students' health status, so as to improve the efficiency of campus health management. The experimental results show that the accuracy and AUC values of the proposed method are 0.9507 and 0.9525, respectively, which can better enhance the health management capability of the smart campus system.

Acknowledgements

This work is supported by the Research Project Supported by Shanxi Scholarship Council of China named: A Research on Digital Technology Enabling the High-quality Development of the National Fitness Public Service System in Shanxi Province (No. 2024-119).

Declarations

All authors declare that they have no conflicts of interest.

References

Al-Tameemi, I.S., Feizi-Derakhshi, M-R., Pashazadeh, S. and Asadpour, M. (2023) 'Multi-model fusion framework using deep learning for visual-textual sentiment classification', *Computers, Materials & Continua*, Vol. 76, No. 2, pp.2145–2177.

Anagnostopoulos, T., Kostakos, P., Zaslavsky, A., Kantzavelou, I., Tsotsolas, N., Salmon, I., Morley, J. and Harle, R. (2021) 'Challenges and solutions of surveillance systems in IoT-enabled smart campus: a survey', *IEEE Access*, Vol. 9, pp.131926–131954.

Cai, W. (2023) 'A model for digital education management information system using wireless communication and BP neural networks', *Mobile Networks and Applications*, Vol. 28, No. 6, pp.2149–2161.

Chao, Z., Yi, L., Min, L. and Long, Y.Y. (2024) 'IoT-enabled prediction model for health monitoring of college students in sports using big data analytics and convolutional neural network', *Mobile Networks and Applications*, Vol. 14, No. 5, pp.1–18.

Dong, Z.Y., Zhang, Y., Yip, C., Swift, S. and Beswick, K. (2020) 'Smart campus: definition, framework, technologies, and services', *IET Smart Cities*, Vol. 2, No. 1, pp.43–54.

Fang, M., Peng, S., Liang, Y., Hung, C-C. and Liu, S. (2023) 'A multimodal fusion model with multi-level attention mechanism for depression detection', *Biomedical Signal Processing and Control*, Vol. 82, p.104561.

Haleem, M.S., Ekuban, A., Antonini, A., Pagliara, S., Pecchia, L. and Allocca, C. (2023) 'Deep-learning-driven techniques for real-time multimodal health and physical data synthesis', *Electronics*, Vol. 12, No. 9, p.1989.

Huang, Y., Yi, J. and Yin, Y. (2024) 'Design and implementation of a multimodal data collection and distribution system based on Python-taking school management system as an example', *Applied and Computational Engineering*, Vol. 120, pp.1–10.

Jafari, A., Ganesan, A., Thalisetty, C.S.K., Sivasubramanian, V., Oates, T. and Mohsenin, T. (2018) 'Sensornet: a scalable and low-power deep convolutional neural network for multimodal data classification', *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 66, No. 1, pp.274–287.

John, A., Redmond, S.J., Cardiff, B. and John, D. (2021) 'A multimodal data fusion technique for heartbeat detection in wearable IoT sensors', *IEEE Internet of Things Journal*, Vol. 9, No. 3, pp.2071–2082.

Kuo, C-C.J. (2016) 'Understanding convolutional neural networks with a mathematical model', *Journal of Visual Communication and Image Representation*, Vol. 41, pp.406–413.

Li, W. (2021) 'Design of smart campus management system based on internet of things technology', *Journal of Intelligent & Fuzzy Systems*, Vol. 40, No. 2, pp.3159–3168.

Li, Z., Liu, F., Yang, W., Peng, S. and Zhou, J. (2021) 'A survey of convolutional neural networks: analysis, applications, and prospects', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 12, pp.6999–7019.

Liang, J., Qin, Z., Xue, L., Lin, X. and Shen, X. (2021) 'Verifiable and secure SVM classification for cloud-based health monitoring services', *IEEE Internet of Things Journal*, Vol. 8, No. 23, pp.17029–17042.

Liang, Y. and Chen, Z. (2018) 'Intelligent and real-time data acquisition for medical monitoring in smart campus', *IEEE Access*, Vol. 6, pp.74836–74846.

Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X. and Zheng, W. (2023) 'The multi-modal fusion in visual question answering: a review of attention mechanisms', *PeerJ Computer Science*, Vol. 9, p.e1400.

Nemati, S., Rohani, R., Basiri, M.E., Abdar, M., Yen, N.Y. and Makarenkov, V. (2019) 'A hybrid latent space data fusion method for multimodal emotion recognition', *IEEE Access*, Vol. 7, pp.172948–172964.

Roda-Sanchez, L., Cirillo, F., Solmaz, G., Jacobs, T., Garrido-Hidalgo, C., Olivares, T. and Kovacs, E. (2023) 'Building a smart campus digital twin: system, analytics, and lessons learned from a real-world project', *IEEE Internet of Things Journal*, Vol. 11, No. 3, pp.4614–4627.

Song, J., Chen, H., Li, C. and Xie, K. (2023) 'MIFM: multimodal information fusion model for educational exercises', *Electronics*, Vol. 12, No. 18, p.3909.

Wang, K-C., Pan, H-W. and Wu, C-E. (2024) 'Smart campus innovative learning model for social practitioners of universities' third mission: to promote good health and well-being', *Sustainability*, Vol. 16, No. 14, p.6017.

Wang, L. (2024) 'Deep learning-based depression analysis among college students using multi modal techniques', *International Journal of Advanced Computer Science & Applications*, Vol. 15, No. 7, p.942.

Xie, Y., Zhan, N., Zhu, Q., Zhan, J., Guo, Z., Qiao, C., Zhu, J. and Xu, B. (2024) 'Multimodal data visualization method for digital twin campus construction', *International Journal of Digital Earth*, Vol. 17, No. 1, p.2431624.

Ye, M., Ruiwen, N., Chang, Z., He, G., Tianli, H., Shijun, L., Yu, S., Tong, Z. and Ying, G. (2021) 'A lightweight model of VGG-16 for remote sensing image classification', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 14, pp.6916–6922.