Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **2**, May **2025**

Multimodal Data Fusion Based on Generative Adversarial Network in University Teaching Evaluation

Min Liu*

School of Computer Science and Software Engineering
Southwest Petroleum University, Nanchong 637001, P. R. China
miner_liu@126.com

Hua Teng

Department of Computer
China West Normal University, Nanchong 637009, P. R. China
398167475@qq.com

Hong Zhang

Faculty of Computer and Information Science
SouthWest University, Chongqing 400700, P. R. China
zhangh@swu.edu.cn

Hao Li

Krirk University, Bangkok 10220, Thailand

*Corresponding author: Min Liu
Received April 16, 2024, revised August 23, 2024, accepted December 16, 2024.

Abstract. *With the advancement of educational evaluation reform and the popularization of higher education, university teaching quality evaluation has been a research hotspot in the field of educational evaluation. In response to the issue of multimodal heterogeneity gap in current university education evaluation data, which makes it tough to integrate evaluation data, this article suggests a Multimodal Data Fusion (MDF) method for university teaching evaluation relied on Generative Adversarial Network (GAN). Firstly, the MDF model is constructed from two aspects: establishing similarity network and designing regularization terms. By designing sparse regularization terms, the feature selection of multimodal samples is carried out to address the issue of large deviation of fusion data. Secondly, BERT pre-training model is adopted to extract the modal embedding features of teaching evaluation text, and Bi-LSTM is adopted to extract the audio and visual modal embedding features of teaching evaluation. Then, on the ground of the constructed MDF model, combined with the GAN and autoencoder, a multi-modal adversarial autoencoder learning model is established to reduce the heterogeneous gap between modes and form a unified representation. At last, the self-attention pretest network is established to achieve the automatic classification of multi-modal university teaching evaluation data. The experimental outcome implies that the accuracy, F1 value and determination coefficient R^2 of the suggested method are 92.18%, 93.72% and 90.5%, respectively, which are superior to the comparison method, showing good classification performance.*
Keywords: University teaching evaluation; Generative adversarial network; Multimodal data fusion; BERT model; Self-attention mechanism

1. **Introduction.** Teaching evaluation is an indispensable part of university teaching, and its goal is to enhance the quality of teaching. On the one hand, it allows students to gain self-knowledge through evaluation, discover their own strengths and weaknesses, and then check and make up for the deficiencies; on the other hand, it allows teachers to have more reflection on teaching through evaluation, and lays the foundation for improving the quality of teaching [1, 2]. Recently, as artificial intelligence technology rapidly growing, the university teaching reform has also undergone great changes, the university teaching mode is more and more multimodal, and this new teaching mode change will inevitably bring about the reform of teaching evaluation [3]. The traditional assessment model is mainly based on summative assessment, which pays too much attention to the memorization of language knowledge and written scores, ignores the learners' enthusiasm and initiative, and is not conducive to a comprehensive and objective assessment of the learners [4, 5]. It can be seen that the traditional university teaching evaluation mode can no longer meet the requirements of modern teaching mode. Multimodal teaching makes up for the shortcomings of traditional teaching evaluation and provides multiple forms for it, so that teaching evaluation can become more objective, scientific and comprehensive [6].

1.1. **Related work.** Moubayed et al. [7] designed a learning performance assessment system by incorporating K-Means clustering algorithm. Goos and Salomons [8] used partial correlation analysis and factor analysis to analyze the indicators affecting the quality of teachers' teaching, and used multiple linear regression to find out the valuable patterns of indicators. Ifenthaler and Widanapathirana [9] combined hierarchical analysis and support vector machine to conduct fuzzy comprehensive evaluation of teaching quality to enhance the scientific results. Furthermore, related studies on the introduction of machine learning technology into university teaching evaluation include: the use of rough set theory to deal with the issue of unreasonable indicator weights [10], the introduction of decision trees to analyze multimodal evaluation data [11], the use of association rule algorithms to analyze factors affecting the quality of teaching [12], and so on.

Currently, as deep learning technology rapidly developing, a variety of teaching evaluation methods relied on deep learning have appeared. Yang et al. [13] relied on optimized BP neural network to quantify the teaching indexes comprehensively, and obtained more reasonable evaluation outcome. Zhang et al. [14] suggested a multimodal teaching evaluation fusion method based on Convolutional Neural Network (CNN). Ge et al. [15] on the ground of deep CNN to reveal the hidden correlation and diversity of multimodal features in teaching evaluation texts and images. Hutchings et al. [16] designed a unified fusion model through feature extraction and information measurement to adaptively integrate the multimodal features. Qianyi and Zhiqiang [17] fused multimodal data using an LSTM-based and an attentional mechanism that appropriate weights are generated to combine features from different modalities relied on the context of the teaching evaluation text. Jeong and Cho [18] designed a Recurrent Neural Network (RNN) to compute displacement vectors by using a gating mechanism to learn the nonlinear combination between the teaching evaluation text and images. This limits the performance of feature fusion of instructional evaluation texts and images due to the large differences in feature distributions between modalities.

To address the issue of feature distribution differences between multimodal instructional evaluation data and to include more information in instructional evaluations, researchers have turned their attention to Generative Adversarial Network (GAN)-based instructional evaluation methods. Sarwat et al. [19] implemented the character-to-pixel conversion process of instructional evaluation texts through GANs which maximized the similarity between the source and the target codes of each category. Aguilar et al. [20] adopted two

discriminative models for both intra-modal and inter-modal discrimination of instructional evaluation text to make the generated common representations more discriminative. Chui et al. [21] suggested a two-discriminator GAN model, which allows the generator to be more adequately trained to maintain similarity with multiple instructional evaluation distributions.

1.2. **Contribution.** On the ground of the above analysis, the current university teaching evaluation methods ignore the heterogeneity gap of multimodal data and the integrity of semantic information. Thus, how to obtain modality-invariant representations in the feature mapping process while reducing the loss of modality-specific representations requires further research. In this article, a university teaching evaluation method based on GAN for MDF is suggested. Firstly, the similarity relationship between each modal sample is considered, and a similarity network is built and a regularization term is designed to optimize the MDF model. Second, the textual modal embedding features of teaching evaluation are extracted using BERT pre-training model, and the audio and visual modal embedding features of teaching evaluation are extracted using Bi-LSTM. Then, relied on the optimized MDF model and GAN, an adversarial encoder-discriminator is built to reduce the modal gaps, so as to obtain the modal invariant representation. Again, a new decoder structure is designed to reduce the loss of modal information. Finally, a self-attentive prediction network is built to realize automatic classification of multimodal university teaching evaluation data. The experimental outcome show that the suggested method has high accuracy, F1 value, coefficient of determination (R^2), and low Mean Absolute Error (MAE) and Mean Square Error (MSE), which verifies the efficiency of the designed method.

2. Theoretical analysis.

2.1. **Generative adversarial network.** At the heart of GAN is the adversarial learning process between a generative network G and a discriminative network D. Specifically, G generates false samples $G(z)$ from a random vector z, while D receives real instance x and false instance $G(z)$ as inputs and outputs probability values to indicate whether the inputs are from the true distribution [22]. This process can be understood as the discriminative network tries to improve its accuracy by constantly distinguishing between real and generated samples, while the generative network deceives the discriminative network by constantly adjusting its parameters to generate more realistic fake instances. This adversarial learning process allows GAN to excel in generated various types of data, which is implied in Figure 1.

Let z denotes the input network, x represents the actual instance, $G(z)$ denote the output of the generator, and the output of the discriminator is divided into $D(G(z))$ and $D(x)$ according to the different types of inputs. Comparing the value of $D(x)$ in the ideal state is 1, and the value of $D(G(z))$ is 0, the adversarial learning process is represented by the following objective function.

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[\log D(x)] + E_{\tilde{z}\sim P_g(\tilde{z})}\left[\log\left(1 - D(G(z))\right)\right] \tag{1}$$

where $\min_G V(D,G)$ is the generator objective function, $V(D,G)$ is the objective function of GAN and $\max_D V(D,G)$ is the discriminator objective function.

By training the GAN network to minimize the difference between $P_g(x,\delta)$ and $P_{data}(x)$, the following likelihood function is established.

$$L = \prod_{i=1}^{n} P_g\left(x^i, \delta\right) \tag{2}$$

where $P_{data}(x)$ denotes the sample-fixed data distribution, $P_g(x, \delta)$ denotes the generated sample distribution, and δ denotes the probability parameter.

The L-likelihood function is maximized to ensure that the difference between $P_g(x, \delta)$ and $P_{data}(x)$ is kept to a minimum, and the optimal δ^* is calculated on this basis.

$$\begin{aligned} \delta^* &= \arg\max \prod_{i=1}^{n} P_g\left(x^i, \delta\right) \\ &= \arg\max_{\delta} \left(\int_x P_{data} \log P_g(x, \delta) dx - \int_x P_{data} \log P_{data}(x) dx \right) \end{aligned} \tag{3}$$

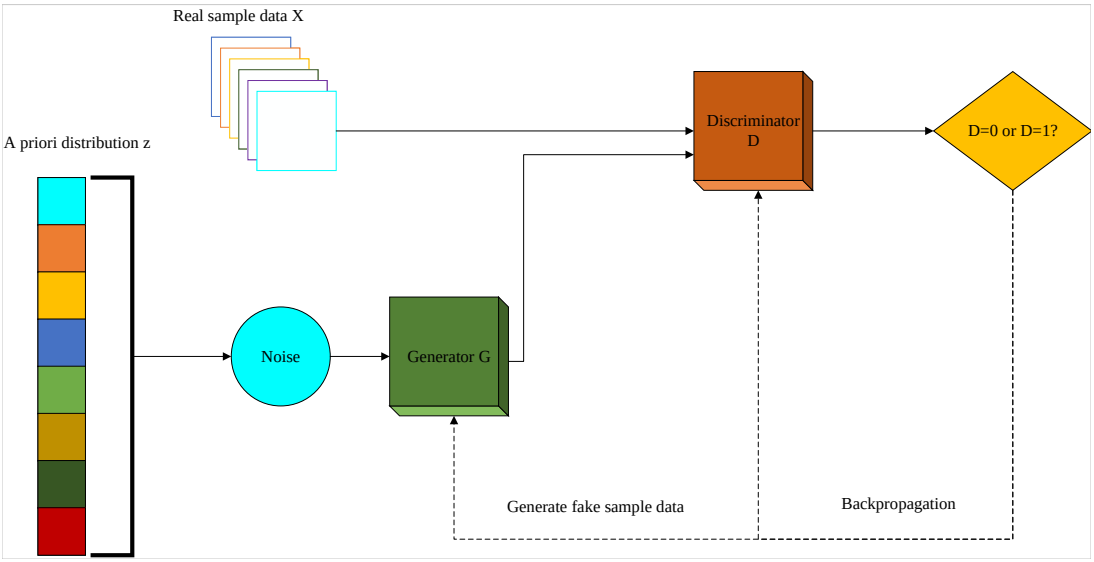

Figure 1. The structure of the GAN

2.2. **Multimodal data fusion.** When a person makes a judgment on a target object, he or she usually obtains different feature information of the target object from different sensory organs of the body [23], and then processes and analyzes various feature information according to the existing a priori knowledge, and finally makes a corresponding judgment according to the results of the analysis. Similarly, MDF also utilizes data from multiple sources, such as text, image, audio, and video, to extract different feature information from them, and then process the information by integrating and removing redundancy, to ultimately arrive at a comprehensive and accurate outcome. The schematic diagram of MDF is implied in Figure 2.

MDF is a method of integrating and processing data from different modalities to improve system performance and accuracy [24]. In this process, redundant information can be used to increase the accuracy of the system through cross-validation of multimodal data, while complementary information can be used to improve the stability of the system through the mutual complementation of data from different modalities.

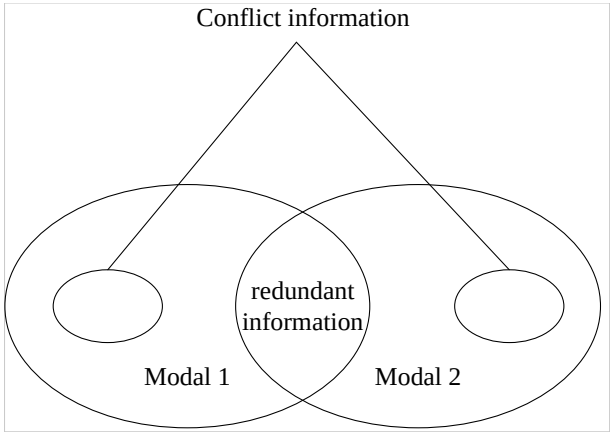

Figure 2. The schematic diagram of MDF

3. **Construction of a multimodal data fusion model.** Current MDF algorithms ignore the structural information of the data, which leads to large bias. It is expected that similar sample objects should have similar response values, so it is necessary to consider the similarity relationship between the modal samples and design sparse regularization terms to choose the features of the samples of multiple modalities. This article will construct a MDF model from the establishment of similarity network and the design of regularization terms.

(1) Similarity network construction. In order to make similar samples in multimodal data have similar response values, the similarity relationship description of Equation (4) is established as bellow.

$$S_{des} = \frac{1}{2}\sum_{i,j}^{L} S_{i,j}^{(n)} \left(\hat{y}^{(n)}(i) - \hat{y}^{(n)}(j)\right)^2 \tag{4}$$

where $\hat{y}^{(n)} = X^{(n)}v^{(n)} \in \mathbb{R}^L$ is the estimated response vector, L is the sample, $\hat{y}^{(n)}$ denotes the label estimation vector for the n-th modality, $\hat{y}^{(n)} \in \mathbb{R}^l$, $X^{(n)}$ denotes the dataset for the n-th modality, $v^{(n)}$ denotes the weight vector for the n-th modality, and $S^{(n)} = \left[S_{i,j}^{(n)}\right] \in \mathbb{R}^{L\times L}$ is defined as the similarity matrix between each pair of samples under the n-th modality.

$$S_{i,j}^{(n)} = \begin{cases} \exp\left(-\frac{\|x_i^n - x_j^n\|_2^2}{\}^{(n)}}\right), & x_i \sim x_j \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\}^{(n)}$ is a free parameter determined by a Gaussian kernel.

Using Laplace operator matrix diagram, Equation (4) is then rewritten as Equation (6) by computing

$$L^{(n)} = D^{(n)} - S^{(n)}, \quad D^{(n)} = diag(d_i), \quad d_i = \sum_{j=1}^{L} S_{i,j}^{(n)}, \quad (1 \le i \le L),$$

where v represents the weight vector [25].

$$S_{des} = \left(\hat{y}^{(n)}\right)^T L^{(n)}\hat{y}^{(n)} = \left(X^{(n)}v^{(n)}\right)^T L^{(n)}\left(X^{(n)}v^{(n)}\right) \tag{6}$$

Based on the regularization constraint term of Equation (6), the MDF model that takes into account the sample structure relationship within the modality is implied in Equation (7).

$$\min_{v} \frac{1}{2} \sum_{n=1}^{N} \|Y - X^{(n)} v^{(n)}\|_2^2 + \sum_{n=1}^{N} \mu_n \|v^{(n)}\|_1 + \sigma \sum_{n=1}^{N} \left(X^{(n)} v^{(n)}\right)^T L^{(n)} \left(X^{(n)} v^{(n)}\right) \quad (7)$$

where A represents the L_1 regularization factor for the n-th mode, and σ denotes the regularization factor.

(2) Regularization term design. The above models do not pay attention to the important correlations between modes and modalities. Therefore, an optimized MDF model is suggested by designing a new regularization term that takes full advantage of the potential relationships existing between different modal data samples. To incorporate the similar relationship between modes into the model of MDF, firstly, the new regularization term is designed, as implied in Equation (8).

$$\varphi(v, \alpha, \beta) = \begin{cases} \alpha \frac{1}{2} \sum_{a,b_i,j}^{N} \sum_{i,j}^{L} S_{i,j}^{(a,a)} \left(\hat{y}^{(a)}(i) - \hat{y}^{(a)}(j)\right)^2, & a = b \\ \beta \frac{1}{2} \sum_{a,b_i,j}^{N} \sum_{i,j}^{L} S_{i,j}^{(a,b)} \left(\hat{y}^{(a)}(i) - \hat{y}^{(b)}(j)\right)^2, & a \neq b \end{cases} \quad (8)$$

where $a = b$ represents the correlation between samples in a single modality, and $a \neq b$ denotes the correlation between samples between modalities, and $S_{i,j}^{(a,b)}$ denotes the similarity between the i-th sample of modality a and the j-th sample of modality b. The two regularization parameters α and β constrain the two corresponding parts, respectively.

Then the similarity between samples within modes and the similarity Laplace matrix L of samples between modes are obtained by Laplace transform as implied in Equation (9).

$$L = \begin{bmatrix} L^{(1,1)} & \cdots \\ \vdots & L^{(N,N)} \end{bmatrix} \quad (9)$$

where the diagonal portion is the Laplace matrix of the samples within each group of modes L, and the other portion off the diagonal is the Laplace matrix of the samples between modes L. After obtaining the structural relationships within and between the modes, the following new regularization term can be obtained.

$$\varphi(v, \alpha, \beta) = \begin{bmatrix} X^{(1)} v^{(1)} \\ \vdots \\ X^{(N)} v^{(N)} \end{bmatrix}^T L \begin{bmatrix} X^{(1)} v^{(1)} \\ \vdots \\ X^{(N)} v^{(N)} \end{bmatrix} \quad (10)$$

The new regularization constraint terms mentioned above are added to the original model to finally derive a new MDF model, as implied in Equation (11) below.

$$H(v) = \min_{v} \frac{1}{2} \sum_{n=1}^{N} \|Y - X^{(n)} v^{(n)}\|_2^2 + \sum_{n=1}^{N} \mu_n \|v^{(n)}\|_1 + \varphi(v, \alpha, \beta) \quad (11)$$

where μ, α and β are the control parameters of the regularizer. The first term is the loss function term. The second term is the L_1 norm regularization term, which aims to sparse the features of each mode. The last term is the prevalence regularization term, which preserves the structural relationship of the samples within modes.

4. **Multimodal data fusion based on generative adversarial networks in university teaching evaluation.**

4.1. **Multimodal feature embedding for teaching evaluation data.** Intending to the issue that modality-specific representations of university teaching evaluation data are ignored and the contribution of multimodal data is hard to be determined, a MDF method relied on GAN is suggested for university teaching evaluation. Firstly, the embedding features of textual, audio and visual modalities of teaching evaluation are extracted using BERT model and Bi-LSTM respectively. Secondly, combining the GAN and the self-encoder, an adversarial encoder-regressor is built to reduce the modal gaps, so as to achieve the modal invariant representation. Third, a new decoder structure is designed to minimize the loss of modal information. Finally, a MDF model is used to fuse modal-invariant and modal-specific representations to gain automatic classification of multimodal teaching evaluation data. The offered method's workflow is implied in Figure 3.

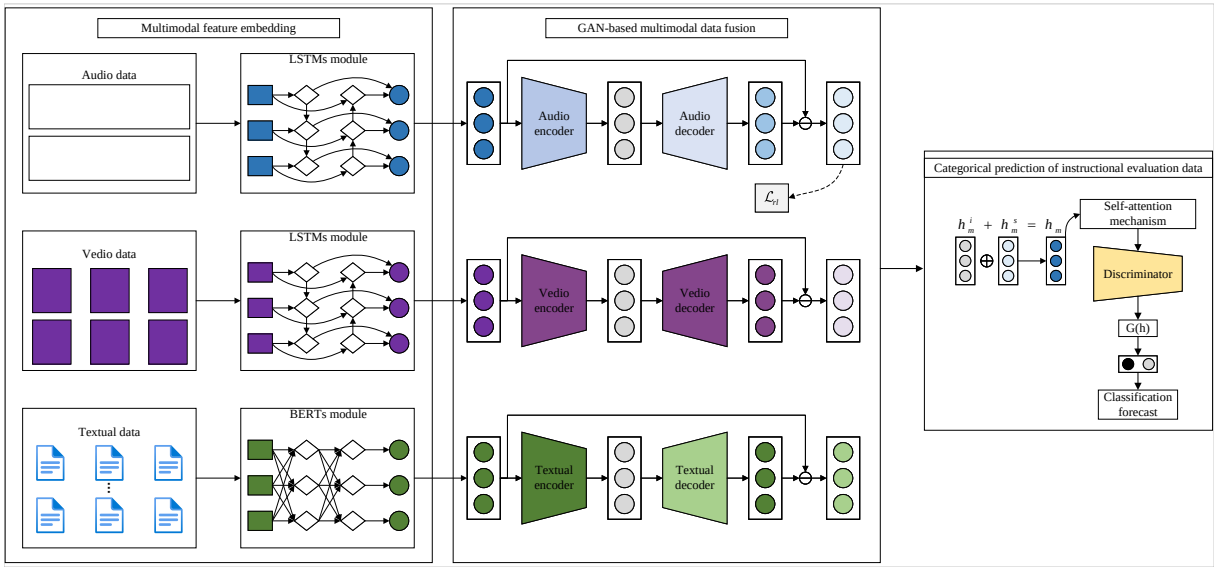

Figure 3. The entire model of the suggested method

Suppose the multimodal instructional evaluation dataset is divided into M segments, each of which contains feature sequences for audio, visual and textual modalities, denoted as $X_b \in R^{s_b \times d_b}$, $X_w \in R^{s_v \times d_v}$ and $X_k \in R^{s_k \times d_k}$, respectively, where $s_n, d_n, n \in \{b, w, k\}$, $d_n, n \in \{b, w, k\}$ denote the length of the feature sequences and the dimensionality of the corresponding features, respectively. Given the above feature sequences, this paper predicts the label y based on GAN and the multimodal data fusion model in the previous section, where y is the score variable ($y \in R$) of the teaching evaluation data.

For audio and visual modes, two LSTM components: $R^{s_n \times d_n} \to R^d, n \in \{b, w\}$ are used to extract features from audio modal data X_b and visual modal data X_w, and the outputs are audio modal feature x_b and visual modal feature x_w. The LSTM module consists of two parts: a bi-directional long- and short-term memory network, Bi-LSTM, and a fully-connected layer, Dense, which is used to extract basic features of the audio modes and visual modes, and the Dense layer is used to receive the hidden layer states from the last moment of Bi-LSTM and unify the different modes dimensionally as indicated below.

$$x_b = LSTM\left(X_b; \delta_b^{lstm}\right) \tag{12}$$

$$x_w = LSTM\left(X_w; \delta_w^{lstm}\right) \tag{13}$$

where δ_b^{lstm} and δ_w^{lstm} denote the parameters of the LSTM components in audio and visual modalities, respectively.

For textual modalities, the gain in model performance is greater than that for non-textual modalities, so it is especially important to learn feature embeddings for textual modalities. This article adopts a pre-trained language model BERT [26] to extract features from textual modal teaching evaluation data. In addition, like LSTM, a Dense layer is connected to the output layer of BERT to form the BERT component: $R^{s_k \times d_k} \to R^d$, and the output is textual modal feature x_k, as indicated below.

$$x_k = BERT\left(X_k; \delta_k^{bert}\right) \tag{14}$$

where δ_k^{bert} is a parameter of the BERT component.

4.2. **GAN-based multimodal data fusion for teaching evaluation.** In this article, textual modalities are adopted as source modalities, audio and visual modalities are used as target modalities, and an adversarial encoder-regressor is adopted to narrow down the distributional differences between the target modalities and the source modalities based on GAN and a designed MDF model.

For each modality $n \in \{b, w, k\}$, an encoder G_n and discriminator D_n are constructed for feature mapping and feature fusion of multimodal features, and multimodal embedded features x_n are mapped as inputs into a shared common subspace.

$$h_n = G_n\left(x_n; \delta_{H_n}\right), n \in \{b, w, k\} \tag{15}$$

where G_n is parameterized by δ_{G_n} as a unimodal encoder and h_n is a modal invariant feature mapped into the common subspace.

G_n consists of a simple fully-connected network that plays two roles simultaneously: as an encoder network in the autocoder and as a generator network in the GAN.

The multimodal features are then discriminated using a discriminator D_n. The features of the textual modality of instructional evaluation are defined as true while the other modalities are defined as false. D_n is to detect the modality of the feature as best as possible given the unknown features of the modality. The task of the encoder is to reduce the heterogeneity between modalities by making the discriminator unable to detect the modality of the feature. The adversarial loss function is defined as bellow.

$$L_b = E_{x_n \sim P_{x_n}}\left[\log\left(D\left(G_k\left(x_k\right)\right)\right) + \log\left(1 - D\left(G_w\left(x_w\right)\right)\right) + \log\left(1 - D\left(G_b\left(x_b\right)\right)\right)\right] \tag{16}$$

Using the loss function defined in the above equation, the modal invariant representation of multimodal features can be learned by iteratively training the encoder and discriminator networks in an adversarial manner. The training process is as bellow.

In the first step, the discriminator network parameters δ_D are updated and the encoder network parameters δ_{G_n} are frozen.

$$J^{(D)}\left(\delta_D, \delta_{G_n}\right) = -\log\left(D\left(G_k\left(x_k\right)\right)\right) - \log\left(1 - D\left(G_w\left(x_w\right)\right)\right) - \log\left(1 - D\left(G_b\left(x_b\right)\right)\right) \tag{17}$$

In the second step, the encoder network parameters δ_{G_n} are updated, the discriminator network parameters δ_D are frozen, and a model is generated aiming to minimize the loss function so that the visual and audio features fit the correlation distributions of the linguistic features as closely as possible.

$$J^{(G_n)}\left(\delta_D, \delta_{G_n}\right) = \log\left(D\left(G_k\left(x_k\right)\right)\right) + \log\left(1 - D\left(G_w\left(x_w\right)\right)\right) + \log\left(1 - D\left(G_b\left(x_b\right)\right)\right) \tag{18}$$

To minimize the risk of information loss during the sign mapping process, this paper uses an encoder-decoder network structure to fuse modal invariant features in the common

subspace. For each unimodal $n \in \{b, w, k\}$: the decoder is constructed separately for fusing feature h_n. The reconstruction loss function is implied below.

$$\hat{x}_n = D_n\left(h_n; \delta_{D_n}\right) \tag{19}$$

$$L_{rk} = \frac{1}{3} \sum_{n \in \{b,w,k\}} \|x_n - \hat{x}_n\|_2^2 \tag{20}$$

4.3. **Categorical prediction of instructional evaluation data.** This article adopts the self-attention mechanism [27] to fuse prediction networks to forecast evaluation score labels by learning multimodal features of evaluation data. To extract more useful information from the multimodal feature values, the self-attention mechanism is applied to the h_n.

$$V_b = \frac{e^{\left(h_n^T \cdot h_n\right)}}{\sum_{i=1}^{M} e^{\left(h_n^T \cdot h_{ni}\right)}}, \quad n \in \{b, w, k\} \tag{21}$$

$$h_n^b = V_b \cdot h_n \tag{22}$$

$$h = concat\left(h_b^b, h_w^b, h_k^b\right) \tag{23}$$

where h_n^T is the transpose vector of h_n, V_b is the self-attention weight, h_n^b is the self-attention guided multimodal feature, and h is the multimodal fusion feature obtained by splicing.

Finally, the multimodal fusion feature h is passed to a fully connected network $\hat{y} = G(h; \delta^G)$ for classification prediction and the classification loss is computed.

$$L = \sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2^2 \tag{24}$$

The total loss of the model is weighted by task loss, reconstruction loss, within-modality contrast loss and cross-modality contrast loss.

$$L_{total} = L + \varphi L_b + \mu L_{rk} \tag{25}$$

where φ denotes the regularization constraint term and μ denotes the regularization parameter.

5. **Performance testing and analysis.**

5.1. **Recognition performance analysis.** To estimate the performance of the suggested university teaching evaluation method, this article experiments randomly sampled university classroom teaching evaluation data from a university's teaching management system to conduct simulation experiments, which contains 11,392 student classroom teaching evaluation data from six semesters as the initial dataset for the experiments. In this article, the teaching evaluation data samples are divided into training set and validation data set according to the ratio of 8:1:1, and the MSACNN [14], ECURNN [18], ETRGAN [20] and the MDF-GAN method designed in this article are trained and tested respectively.

The experimental platform is configured as Ubuntu 23.04 LTS Linux operating system, Intel(R) Core(TM) i5-13500H CPU @ 2.60GHz, 16 GB RAM, and Python V2.7. A 768-dimensional pre-trained BERT model is adopted to generate word embedding vectors, and the batch size is set to 32. Batch-size is set to 32, and the number of layers of the neural network is set to 4. The Adam optimizer is adopted to optimize the parameters of

the model during the training process. In addition, to avoid overfitting, the learning rate and dropout rate are set to 0.0001 and 0.1, respectively.

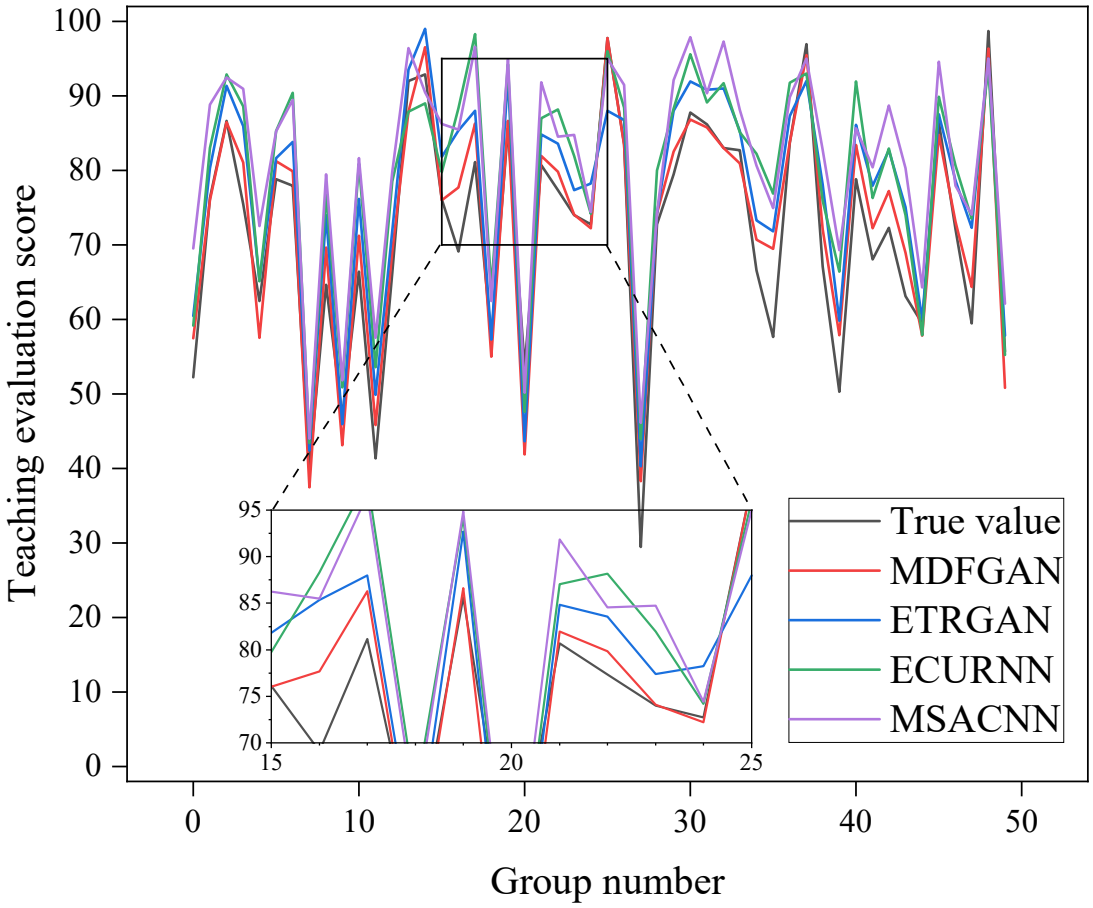

Figure 4. Comparison of predictive scores for different teaching evaluation methods

For the purpose of reflecting the effectiveness of MDFGAN method more realistically, this article selects 50 groups of teaching evaluation scores of the same teacher of the same course in a university, and compares the outcome of the predicted scores of MSACNN, ECURNN, ETRGAN and MDFGAN methods with the real data, and the outcome is indicated in Figure 4. It can be seen that the MDFGAN model has a very small deviation and the prediction curve is close to the real value, while the ETRGAN method model can better predict the trend of the real value and has a smaller prediction deviation, and most of the experimental results of the ECURNN model are far away from the error-tolerance interval. The MSACNN model only uses the CNN method to extract the features of the indexes that affect the text of the teaching evaluation and does not capture the multimodal features of the data of these indexes, resulting in the prediction results being far away from the error-tolerance interval. In summary, the prediction results of the MDFGAN method are better than those of MSACNN, ECURNN and ETRGAN.

This article conducts comparative experiments of different teaching evaluation methods using a combination of Accuracy (Acc), the reconciled value of precision and recall F1 [28], the mean absolute error (MAE), the mean square error (MSE), and the coefficient of determination (R^2) [29] performance metrics. Table 1 demonstrates the results of comparative experiments of different methods.

Table 1. Comparison of the categorization performance of various evaluation methods

Method	**Acc(%)**	**F1(%)**	**MAE(%)**	**MSE(%)**	$\mathbf{R^2}$(%)
MSACNN	76.44	75.22	19.25	21.54	75.91
ECURNN	80.90	78.16	16.42	17.83	80.46
ETRGAN	87.35	86.98	9.31	11.69	85.17
MDFGAN	92.18	93.72	4.63	5.86	90.50

The analysis reveals that the MDFGAN model improves in performance metrics with 92.18%, 93.72%, 4.63%, 5.86%, and 90.5% on Acc, F1, MAE, MSE, and R^2, respectively, which are superior to those of other comparison methods, indicating the effectiveness of the proposed method.

when compared to the three instructional evaluation methods, namely, MSACNN, ECURNN, and ETRGAN. Particularly, 20.59%, 13.92% and 5.53% are improved on Acc, 24.59%, 19.91% and 7.75% on F1 and 19.22%, 12.48% and 6.26% on R^2, respectively.

This is because the MSACNN method is relied on CNN to extract and classify the text features of teaching evaluation, but it needs to fix the size of the convolutional kernel window, which is not useful when facing longer teaching evaluation sequence information. The ECURNN method utilizes RNN to capture the unimodal features of the teaching evaluation data, but does not extract the multimodal features of the evaluation data, which leads to the general classification effect. The ETRGAN method utilizes two discriminative models of GAN to simultaneously perform intra-modal and inter-modal discrimination of teaching evaluation text, without considering the inter-modal heterogeneity. The method MDFGAN adopts the framework of multimodal adversarial self-encoder based on GAN before fusion of multimodal data, which potentially maps the features of different modalities into a common subspace and reduces the heterogeneity gap between modalities, so the classification efficiency is higher.

5.2. **Ablation experiment.** To better validate the impact of the GAN module and the MDF module in MDFGAN, two comparative models are designed for the analysis, in which the method of removing the GAN module is denoted as MDFGAN/GAN, and the method of removing the MDF module is denoted as MDFGAN/MDF. The outcome of the four models' evaluation metrics are given in Table 2, and the results are plotted on the visual bar charts for the comparative outcome, which are implied in Figure 5.

Table 2. Comparison of experimental results of ablation of components

Method	**Acc (%)**	**F1 (%)**	**MAE (%)**	**MSE (%)**	$\mathbf{R^2}$(%)
MDFGAN/GAN	79.42	80.39	11.48	13.95	77.16
MDFGAN/MDF	73.76	71.64	15.53	19.28	73.22
MDFGAN	92.18	93.72	4.63	5.86	90.50

As can be seen from Figure 5 and Table 2, the accuracy evaluation indexes of MDFGAN are significantly better than those of MDFGAN/GAN and MDFGAN/MDF, in which the MAE of the MDFGAN method is 4.63%, which is 6.85% and 10.9% lower than those of MDFGAN/GAN and MDFGAN/MDF, respectively; the MSE of the MDFGAN method is 5.86%, which is 8.09% and 13.42% lower than those of MDFGAN/GAN and MDFGAN/MDF, respectively. which is 8.09% and 13.42% lower compared to MDFGAN/GAN and MDFGAN/MDF, respectively. Comparing the coefficient of determination R^2, the R^2 value of the MDFGAN method is 90.5%, which is 13.34% and 17.28% higher compared

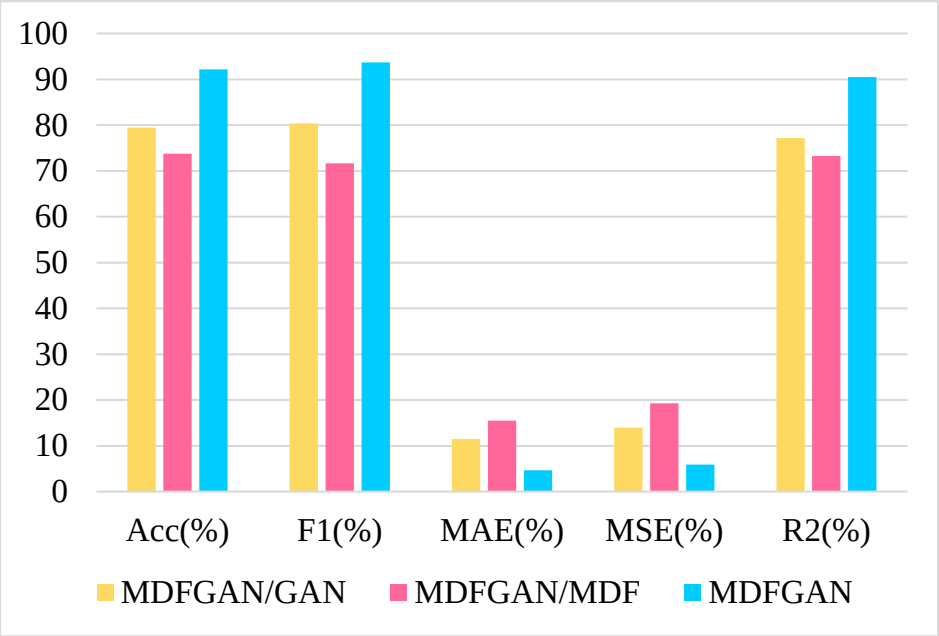

Figure 5. Comparison of performance indexes in ablation experiments

to MDFGAN/GAN and MDFGAN/MDF respectively. Through the above analysis, it can be observed that there is a significant decrease in the recommendation performance of MDFGAN after removing the GAN module and the MDF module, which proves the importance of the MDF for GAN-based teaching and learning evaluation, and thus the MDFGAN that fuses all the modules achieves the best performance.

6. **Conclusion.** Current university teaching evaluation methods ignore the heterogeneity gap of multimodal data and the integrity of semantic information, to deal with the above issues, this article investigates a GAN-based MDF method for university teaching evaluation. Firstly, the similarity network is built and the regularization term is designed to optimize the MDF model to solve the problem of large data deviation. Secondly, BERT pre-training model is used to extract textual modal embedding features for teaching evaluation, and Bi-LSTM is used to extract audio and visual modal embedding features for teaching evaluation. Then, relied on the enhanced MDF model, combining the GAN and the self-encoder, the multimodal adversarial self-encoder learning model is established, and the adversarial encoder-discriminative reduces the modal gap, so as to obtain the modal invariant representation. Finally, a self-attentive prediction network is constructed to realize automatic classification of multimodal data for university teaching evaluation. The experimental outcome implies that the suggested method can effectively improve the classification accuracy, F1 value and R^2 of teaching evaluation data.

Acknowledgment. This work is supported by The Ministry of Education's Industry University Cooperation Collaborative Education Project (2023-2025) (No. 230803879190229).

REFERENCES

[1] F. I. Vin-Mbah, "Learning and teaching methodology," *Journal of Educational and Social Research*, vol. 2, no. 4, pp. 111–118, 2012.

[2] A. Gunn, "Metrics and methodologies for measuring teaching quality in higher education: developing the Teaching Excellence Framework (TEF)," *Educational Review*, vol. 70, no. 2, pp. 129–148, 2018.

[3] J. Hu, "Teaching evaluation system by use of machine learning and artificial intelligence methods," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 5, pp. 87–101, 2021.
[4] T. A. Wolfer and M. M. Johnson, "Re-evaluating student evaluation of teaching: The teaching evaluation form," *Journal of Social Work Education*, vol. 39, no. 1, pp. 111–121, 2003.
[5] L. Zhao, P. Xu, Y. Chen, and S. Yan, "A literature review of the research on students' evaluation of teaching in higher education," *Frontiers in Psychology*, vol. 13, 1004487, 2022.
[6] L. Tan, K. Zammit, J. D'warte, and A. Gearside, "Assessing multimodal literacies in practice: A critical review of its implementations in educational settings," *Language and Education*, vol. 34, no. 2, pp. 97–114, 2020.
[7] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student engagement level in an e-learning environment: Clustering using k-means," *American Journal of Distance Education*, vol. 34, no. 2, pp. 137–156, 2020.
[8] M. Goos and A. Salomons, "Measuring teaching quality in higher education: assessing selection bias in course evaluations," *Research in Higher Education*, vol. 58, pp. 341–364, 2017.
[9] D. Ifenthaler and C. Widanapathirana, "Development and validation of a learning analytics framework: Two case studies using support vector machines," *Technology, Knowledge and Learning*, vol. 19, pp. 221–240, 2014.
[10] C. H. Cheng, L. Y. Wei, and Y. H. Chen, "A new e-learning achievement evaluation model based on rough set and similarity filter," *Computational Intelligence*, vol. 27, no. 2, pp. 260–279, 2011.
[11] E. Park and J. Dooris, "Predicting student evaluations of teaching using decision tree analysis," *Assessment & Evaluation in Higher Education*, vol. 45, no. 5, pp. 776–793, 2020.
[12] Z. Wang, Q. Tian, and X. Duan, "Research on the evaluation index system of college students' class teaching quality based on association algorithm," *Cluster Computing*, vol. 22, no. 6, pp. 13797–13803, 2019.
[13] X. Yang, J. Zhou, and D. Wen, "An optimized BP neural network model for teaching management evaluation," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 3215–3221, 2021.
[14] Y. Zhang, R. An, S. Liu, J. Cui, and X. Shang, "Predicting and understanding student learning performance using multi-source sparse attention convolutional neural networks," *IEEE Transactions on Big Data*, vol. 9, no. 1, pp. 118–132, 2021.
[15] D. Ge, X. Wang, and J. Liu, "A teaching quality evaluation model for preschool teachers based on deep learning," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 3, pp. 127–143, 2021.
[16] P. Hutchings, M. T. Huber, and A. Ciccone, "Feature essays: Getting there: An integrative vision of the scholarship of teaching and learning," *International Journal for the Scholarship of Teaching and Learning*, vol. 5, no. 1, 31, 2011.
[17] Z. Qianyi and L. Zhiqiang, "Research on multimodal based learning evaluation method in smart classroom," *Learning and Motivation*, vol. 84, 101943, 2023.
[18] Y.-S. Jeong and N.-W. Cho, "Evaluation of e-learners' concentration using recurrent neural networks," *The Journal of Supercomputing*, vol. 79, no. 4, pp. 4146–4163, 2023.
[19] S. Sarwat, N. Ullah, S. Sadiq, R. Saleem, M. Umer, A. A. Eshmawi, A. Mohamed, and I. Ashraf, "Predicting students' academic performance with conditional generative adversarial network and deep SVM," *Sensors*, vol. 22, no. 13, 4834, 2022.
[20] A. B. Aguilar, D. Castellanos-Nieves, J. J. Sosa-Alonso, and M. Area-Moreira, "Use of generative adversarial networks (GANs) in educational technology research," *Journal of New Approaches in Educational Research*, vol. 12, no. 1, pp. 153–170, 2023.
[21] K. T. Chui, R. W. Liu, M. Zhao, and P. O. De Pablos, "Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine," *IEEE Access*, vol. 8, pp. 86745–86752, 2020.
[22] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19–46, 2024.
[23] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, 2339, 2023.
[24] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, 1977, 2023.
[25] D. Horak and J. Jost, "Interlacing inequalities for eigenvalues of discrete Laplace operators," *Annals of Global Analysis and Geometry*, vol. 43, pp. 177–207, 2013.

[26] C. Wang, S. Dai, Y. Wang, F. Yang, M. Qiu, K. Chen, W. Zhou, and J. Huang, "Arobert: An asr robust pre-trained language model for spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1207–1218, 2022.

[27] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, 2020.

[28] S. Sergis, D. G. Sampson, M. J. Rodríguez-Triana, D. Gillet, L. Pelliccione, and T. de Jong, "Using educational data from teaching and learning to inform teachers' reflective educational design in inquiry-based STEM education," *Computers in Human Behavior*, vol. 92, pp. 724–738, 2019.

[29] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, e623, 2021.