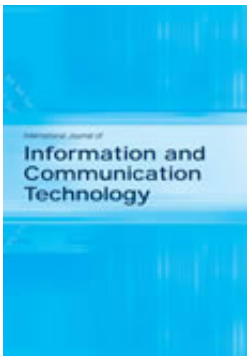
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Dynamic evaluation of college students' psychological state based on multimodal physiological signal fusion and deep generation model

Jing Li

DOI: 10.1504/IJICT.2025.10071345

Article History:

Received:	20 March 2025
Last revised:	10 April 2025
Accepted:	11 April 2025
Published online:	11 June 2025

Dynamic evaluation of college students' psychological state based on multimodal physiological signal fusion and deep generation model

Jing Li

Psychologically Healthy Education Center,
Inner Mongolia University of Finance and Economics,
Huhehot 010000, China
Email: 13314714638@163.com

Abstract: The dynamic assessment of the psychological state of college students is an important research direction in mental health management. In response to the problem of insufficient capture of psychological state changes by existing methods, this paper proposes a dynamic assessment method that combines multimodal physiological signal fusion and deep generation models. Firstly, collect multimodal physiological data and eliminate noise through timing synchronisation and data pre-processing techniques. Secondly, utilising a multimodal feature extraction network based on transformer structure to achieve feature fusion of physiological signals. Subsequently, an improved variational autoencoder (VAE) was designed, combined with an LSTM model, to predict the trend of psychological state changes. Technical support for real-time monitoring and tailored intervention of college students' mental health status is provided by the experimental results showing that the suggested method outperforms current methods in terms of accuracy in psychological state classification and dynamic prediction performance.

Keywords: multimodal physiological signals; dynamic assessment of psychological state; deep generative model; feature fusion; variational autoencoder; VAE.

Reference to this paper should be made as follows: Li, J. (2025) 'Dynamic evaluation of college students' psychological state based on multimodal physiological signal fusion and deep generation model', *Int. J. Information and Communication Technology*, Vol. 26, No. 17, pp.133–146.

Biographical notes: Jing Li obtained her Master's degree from the Xi'an Jiaotong University in 2007 and currently works at the Mental Health Education Center of Inner Mongolia University of Finance and Economics. Her main research focus is on data mining, mental health education.

1 Introduction

The mental health issues of college students are increasingly becoming a focus of social attention (Rodríguez-Romo et al., 2022). With the acceleration of modern life pace and the increase of academic pressure, many college students have experienced psychological problems such as anxiety and depression (Sheldon et al., 2021). The sustained development of this state may have profound impacts on their academic, social, and future lives. Therefore, how to efficiently and accurately evaluate the psychological state of college students, especially dynamically monitor their mental health changes, has become an important research topic in the fields of psychology, education, and computer science (Kang et al., 2021). However, due to the subjectivity of psychological states and the complexity of their dynamic changes, traditional methods such as questionnaire based psychological assessment often have limitations such as poor real-time performance, low accuracy, and subjective results (Auerbach et al., 2016). In recent years, with the rapid development of multimodal physiological signals and deep learning technology, the dynamic assessment of psychological states is gradually achieving a leap from qualitative to quantitative, especially demonstrating significant advantages in real-time and personalised intervention capabilities (Cosoli et al., 2021).

With great real-time performance and great resilience to subjective interference, multimodal physiological signals, such as heart rate variability (HRV), electrodermal activity (EDA), electroencephalography (EEG), etc. are important objective indicators of psychological status and are therefore widely used in mental health research. HRV is a key indicator of autonomic nervous system activity revealed by analysing time series signals of heart rate changes. Research has shown that HRV can reflect an individual's emotional fluctuations and stress levels. For example, Shaffer and Ginsberg (2017) briefly reviewed current views on the mechanisms of 24-hour, short-term (five-minute), and ultra short term (< five-minute) HRV generation, the importance of HRV, and its impact on health and performance. Posada-Quintero and Chon (2020) pointed out that EDA not only contains information in the slow changes (pitch components) represented by the mean, but also in the fast or phase changes of the signal. In emotion classification research, Liu et al. (2011) focused on identifying 'intrinsic' emotions from electroencephalogram (EEG) signals. A real-time fractal dimension based algorithm was proposed to quantify basic emotions using the arousal valence emotion model. Two emotion induction experiments were proposed and implemented, using music stimulation and sound stimulation from the International Emotionally Digital Sound (IADS) database, respectively. Finally, a real-time algorithm was proposed, implemented, and tested to identify six emotions: fear, frustration, sadness, happiness, joy, and satisfaction. In addition, Baltrušaitis et al. (2018) investigated the latest developments in multimodal machine learning itself and introduced them in a general classification system.

Deep learning technology's strong feature extracting and nonlinear modelling powers have made it extensively used in the field of psychological state evaluation (Liu and Liu, 2021). Aldayel et al. (2020) employed a deep learning approach that utilises EEG signals from the DEAP dataset to detect preferences by considering power spectral density and price features. The results show that although the proposed deep learning algorithm exhibits higher accuracy, recall, and precision compared to k-nearest neighbours and support vector machine algorithms, the random forest achieves similar results to deep learning on the same dataset. Recurrent neural networks (RNNs) and their variants, long short-term memory networks (LSTMs), focus on processing time series data and can

capture long-range dependencies of physiological signals. The spatiotemporal hybrid network combined with CNN further improves the classification accuracy of multimodal psychological states (Chen et al., 2023). Generative adversarial networks (GANs) and variational autoencoders (VAEs) have emerged as deep generative models in psychological state assessment in recent years (Goodfellow et al., 2020).

While the above described research shows great development, the technology of dynamic assessment of psychological states still faces the following difficulties: inadequate depth and efficiency of multimodal feature fusion; present research generally uses simple feature concatenation or decision fusion, which makes it difficult to totally examine the complementarity and temporal correlation between signals. Most studies concentrate on stationary psychological state classification, thereby lacking the capacity to detect and forecast dynamic changes in psychological states. Generative models have limited application depth; their promise in psychological state modelling has not been fully realised, particularly in dynamic generation and time series prediction; so, there is still much need for development of current techniques.

In response to the foregoing problems, this paper suggests a dynamic assessment technique for the psychological status of college students combining deep generation models with multimodal physiological signal fusion. The major gifts are:

1 Propose a multimodal feature extraction network based on Transformer architecture, which efficiently captures the complementarity and temporal correlation between signals.

2 An improved VAE was designed, combined with LSTM to generate dynamic changes in psychological states, which improved the accuracy and robustness of dynamic evaluation.

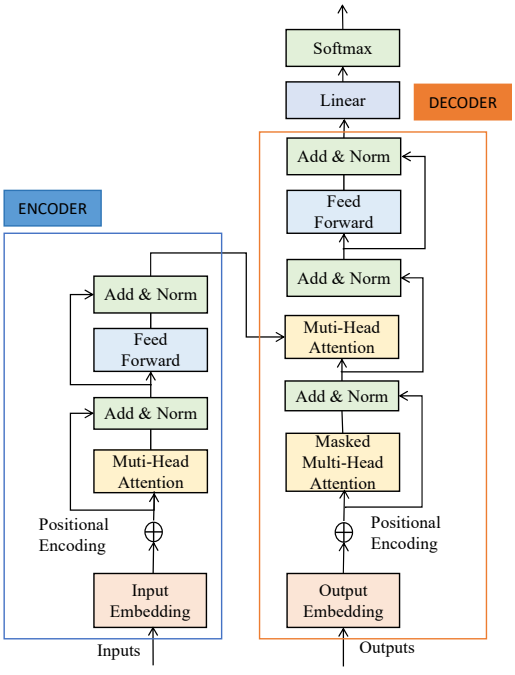
3 Using a dataset of college students' mental health, experimental validation was carried out on which our approach beats current methods in both classification accuracy and dynamic prediction performance, so offering technical support for real-time monitoring and tailored intervention of psychological states.

2 Relevant technologies

2.1 Transformer model

Characterised by a totally self attention based and parallelised architecture design, which abandons the constraints of conventional RNNs or convolutional neural networks in processing sequential data, the transformer model is a revolutionary architecture in the field of deep learning (Frigant and Jullien, 2014). A transformer consists fundamentally in encoder and decoder (Khan et al., 2022). Encoders help to encode input sequences into context sensitive feature representations. The decoder generates the last result depending on the target sequence and encoder output. Every module consists of numerous layered sub layers: multi head self attention technique, feedforward neural network, residual connections, and layer normalisation (Olivares-Galván et al., 2009). Figure 1 shows the model diagram.

Figure 1 Schematic diagram of transformer model (see online version for colours)



The self-attention mechanism seeks to give weights to every sequence position, therefore capturing global dependencies. It computes a weight matrix, then applies weighted summation using each pair of words in the input sequence.

Given input matrix $X \in R^{n\times d}$ (where n is the sequence length and d is the embedding dimension), first project it into a query vector (Q), a key vector (K), and a value vector (V), using the following equation:

$$Q = XW_Q \tag{1}$$

$$K = XW_K \tag{2}$$

$$V = XW_V \tag{3}$$

where $W_Q, W_K, W_V \in R^{n\times d_K}$ is the learnable weight matrix, and d_k is the hidden space dimension of the attention mechanism. Next, calculate the attention score (attention weight):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where $QK^T \in R^{n\times m}$ is the attention score matrix, which represents the importance of each position in the sequence to other positions. $\sqrt{d_k}$ is the scaling factor to avoid gradient instability caused by excessive inner product values. The softmax function normalises scores into probability distributions.

Multihead attention is proposed since single head attention may not be able to capture intricate characteristics. Multiple attention heads of parallel computing improve the expressive capability of the model:

$$MultiHead(Q, K, V) = Concat\left(head_1, ..., head_h\right)W_o \tag{5}$$

The calculation of each attention head is:

$$head_i = Attention\left(QW_Q^i, KW_K^i, VW_V^i\right) \tag{6}$$

where h is the number of attention heads. W_Q^i, W_K^i, W_V^i is the projection matrix of the i^{th} attention head. W_o is the output projection matrix.

Transformers lack a cyclic structure; hence, it is necessary to especially supply positional information to preserve the sequential traits of the input sequence. Position encoding uses a fixed sine function:

$$PE(pos, 2i) = \sin\left(\frac{pos}{1,000^{\frac{2i}{d}}}\right) \tag{7}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{1,000^{\frac{2i}{d}}}\right) \tag{8}$$

where pos is the position of a word in the sequence. i stands for the embedded vector's dimension index. Together with content and spatial information, the produced encoding is included into the input embedding vector.

Following the self attention process in every transformer layer is a feedforward neural network. The equation looks like:

$$FFN(X) = \mathrm{Re}\,LU\left(XW_1 + b_1\right)W_2 + b_2 \tag{9}$$

where W_1 and W_2 are weight matrices.

Residual connection and layer normalisation helps each sub layer produce better outputs for training stability:

$$Output = LayerNorm(X + SubLayer(X)) \tag{10}$$

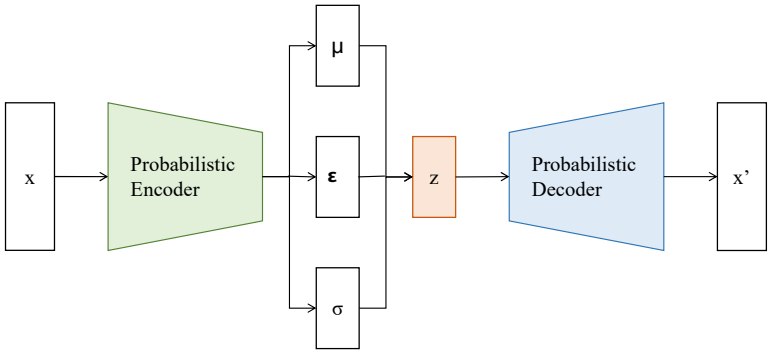
Residual connections solve gradient vanishing in deep network training. By standardising the input distribution, layer normalising speeds convergence.

Transformer's theoretical heart is self attention mechanism and its extension (multi head attention), together with methods including position encoding, feedforward neural network, residual connection, and layer normalisation, so creating a potent and effective sequence modelling framework (Si et al., 2022). With its architecture, transformer has created a fresh paradigm for the field of deep learning with outstanding performance in tasks including natural language processing, time series analysis, and multimodal data fusion.

2.2 Variational autoencoder

VAE is a generative model combining deep learning's benefits with those of probabilistic graph models. It presents probabilistic inference theory to enable better under control and interpretable data creation. VAE is fundamentally based on using deep neural networks to rapidly learn latent variables via variational inference and estimate the latent distribution of high-dimensional data (Vahdat and Kautz, 2020). Figure 2 shows the VAE model; μ and σ respectively reflect the mean and standard deviation of the Gaussian distribution. One may export them from the decoder output. ε can be seen as a kind of random noise applied to preserve z's randomness and produce ε from a normal distribution.

Figure 2 VAE model schematic diagram (see online version for colours)



VAE is based on a probabilistic generative model framework, assuming that the observed data x is generated by the latent variable z. The generation process is modelled as: sampling the latent variable z from the prior distribution $p(z)$, and then generating the observed data x based on the generative distribution $p_\theta(x|\, z)$.

Optimise the model by maximising the edge likelihood $p_\theta(x)$ of the observed data. However, directly optimising $p_\theta(x)$ is usually not feasible because it involves integrating all possible z values:

$$p_\theta(x) = \int p_\theta(x \mid z) p(z) dz \tag{11}$$

In high-dimensional settings, computing this integral can be somewhat challenging. VAE thus developed variational inference to simulate this distribution.

To approximate $p_\theta(x)$, an approximate posterior distribution $q_\theta(x|\, z)$ was introduced, and then the evidence lower bound (ELBO) was maximised instead of directly optimising $p_\theta(x)$. The ELBO equation is:

$$\log p_\theta(x) \geq E_{q_\Phi(z|x)}\left[\log p_\theta(x \mid z)\right] - D_{KL}\left(q_\Phi(x \mid z) \| p(z)\right) \tag{12}$$

Reconstruction error comes first and shows how similar the produced data is to the actual data. The second is Kullback Leibler divergence, which represents the difference between the approximate posterior distribution $q_\Phi(x|\, z)$ and the prior distribution $p(z)$.

Maximising ELBO is equivalent to simultaneously minimising reconstruction error and KL divergence, thereby achieving a distribution $q_\Phi(x|\, z)$ in the latent space that is

close to the prior distribution $p(z)$, and a good representation of data x in the latent space, making it easy to generate.

VAE consists of a decoder and an encoder. The goal of the encoder is to map the observed data x to the distribution parameters (mean and variance) of the latent variable z. Assuming $q_\Phi(x|z)$ is a multidimensional Gaussian distribution:

$$q_\Phi(x \mid z) = N\left(z \middle| \mu_\Phi(x), diag\left(\sigma_\Phi^2(x)\right)\right) \tag{13}$$

The encoder network outputs the mean $\mu_\Phi(x)$ and logarithmic variance $\log \sigma_\Phi^2(x)$ of the latent variables, and the specific process is as follows:

$$\mu_\Phi(x) = f_\Phi^\mu(x), \log \sigma_\Phi^2(x) = f_\Phi^\sigma(x) \tag{14}$$

where f_Φ^μ and f_Φ^σ are neural networks.

The goal of the decoder is to reconstruct data x from latent variable z. The generated distribution $p_\theta(x|z)$ can also be assumed to be a Gaussian distribution:

$$p_\theta(x \mid z) = N\left(x \middle| \mu_\theta(z), I\right) \tag{15}$$

The average of the reconstructed data output by the decoder network is $\mu_\theta(z)$, which is obtained through network mapping z:

$$\mu_\theta(z) = g_\theta(z) \tag{16}$$

Direct sampling from the distribution will cause the gradient to be unable to be backpropagated, therefore sampling the latent variable z. VAE thus embraced a reparameterising trick:

$$z = \mu_\Phi(x) + \sigma_\Phi(x) \odot \varepsilon, \varepsilon \sim N(0, I) \tag{17}$$

where ε is random noise sampled from the standard normal distribution. Through this technique, gradients can optimise network parameters Φ and θ.

VAE is by effectively learning latent spatial structure characteristics from complicated data by aggregating the theories of deep learning and probabilistic graph models (Islam et al., 2021).

2.3 LSTM

LSTM is a unique kind of RNN intended to tackle the problem of vanishing or exploding gradients in standard RNNs when learning long-term dependencies (Greff et al., 2016). By including gating mechanisms and well crafted unit structures, LSTM efficiently captures both short-term and long-term dependencies in sequential data (Yadav and Thakkar, 2024).

Although RNN is one of the primary models for handling sequential input, its chain structure causes gradients to either swiftly erode (vanishing) or endlessly expand (exploding) when backpropagating over extended sequences. Ordinary RNNs find it challenging to recall contextual information from longer time steps as result (Kratzert et al., 2024). LSTM may perform well in long-term dependent sequence learning tasks via design of specific memory cells and gating mechanisms. LSTM's fundamental

concept is to use cell state and gating processes to regulate memory and forgetting of information, therefore addressing long-term reliance.

3 A dynamic evaluation method combining multimodal physiological signal fusion and deep generation model

Combining multimodal physiological signal fusion with deep generative models, this paper suggests a dynamic psychological state evaluation technique. Three essential elements define the approach: dynamic modelling of hidden space using deep generative models, multimodal physiological signal pre-processing and feature extraction, and prediction of psychological state change trends.

3.1 Acquisition and pre-processing of multimodal physiological signals

Included among multimodal physiological signals are HRV, EDA, and EEG. These signals come from electroencephalogram devices, skin conductance measuring tools, and electrocardiogram sensors, in that sequence. Following timing synchronisation methods helps to guarantee signal consistency and timing alignment:

$$t_{aligned} = t_{source} + \Delta t_{calibration} \tag{18}$$

where $t_{aligned}$ represents the aligned timestamp, t_{source} is the original timestamp, and $\Delta t_{calibration}$ is the correction value estimated based on time deviation.

Pre-process the collected signals as follows:

- Denoising: wavelet transform is used for signal denoising. Given signal $x(t)$, its wavelet decomposition is:

$$x(t) = \sum_{j} \sum_{k} c_{j,k} \Psi j, k(t) \tag{19}$$

  where $\Psi j, k(t)$ is the wavelet basis function, and $c_{j,k}$ is the wavelet coefficient.

- Normalisation: normalise the modal signals using the following equation:

$$x_{normalised} = \frac{x - \mu_x}{\sigma_x} \tag{20}$$

where μ_x and σ_x are the mean and standard deviation of signal x, respectively.

3.2 Multi modal feature extraction and fusion based on transformer

Varied modalities of physiological signals have varied temporal traits and information distribution. Every modality is encoded in a transformer-based feature extraction network in order to replicate the characteristics of the signal. The major actions consist in:

- Input embedding: given time series signals of HRV, EDA, and EEG as X_{HRV}, X_{EDA}, and X_{EEG} respectively, generate input embeddings through linear transformation:

$$z_{modality} = W_{embed} X_{modality} + b_{embed} \tag{21}$$

A multi head self attention method is proposed to capture temporal correlations. Key, query, and value of the provided input sequence are Q, K, V. Calculating self-attention requires:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{22}$$

where $\sqrt{d_k}$ is the dimension of the key.

- Feature output: the temporal correlation of multimodal signal features is obtained through self attention mechanism, which are F_{HRV}, F_{EDA}, and F_{EEG}, respectively

By linearly fusing the features of each modality, a multimodal feature representation is obtained:

$$F_{fusion} = \alpha_1 F_{HRV} + \alpha_2 F_{EDA} + \alpha_3 F_{EEG} \tag{23}$$

where α_1, α_2, α_3 is a trainable weight that satisfies $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

3.3 Improved VAE model

In order to generate dynamic features of psychological states, an improved VAE is used to model multimodal feature F_{fusion}. VAE models data generation by introducing probability distributions and its key steps include:

- Encoder: generate distribution parameters of latent variables for input feature F_{fusion}:

$$q_\Phi\left(z \middle| F_{fusion}\right) \sim N\left(\mu, \sigma^2\right) \tag{24}$$

where $\mu = f_u(F_{fusion})$ and $\sigma = f_\sigma(F_{fusion})$ are generated by neural networks.

- Reparameterisation technique: transform the sampling process into differentiable operations through reparameterisation:

$$z = \mu + \sigma \odot \varepsilon, \varepsilon \sim N(0, I) \tag{25}$$

- Decoder: reconstruct features from hidden variable z:

$$p_\theta\left(F_{fusion} \mid z\right) = f_{decode}(z) \tag{26}$$

- Loss function: the loss function consists of reconstruction error and KL divergence:

$$L_{VAE} = E_{q_\Phi(z|F_{fusion})}\left[\log p_\theta\left(F_{fusion} \mid z\right)\right] - KL\left(q_\Phi\left(F_{fusion} \mid z\right) \| p(z)\right) \tag{27}$$

Combining the latent space representation of z, model the trend of psychological state changes through LSTM. Given sequence $\{z_1, z_2, \ldots, z_t\}$, the formula for LSTM recursive update is:

$$h_t = f_{LSTM}\left(z_t, h_{t-1}\right) \tag{28}$$

$$y_t = W_{out} h_t + b_{out} \tag{29}$$

where h_t is the hidden state, and y_t is the predicted psychological state feature. Finally, by combining the dynamic generation of latent space and time series prediction, a dynamic assessment of psychological states can be achieved.

Combining multimodal feature extraction, enhanced VAE generating of latent space distribution, and LSTM time series modelling, this method fully achieves the dynamic evaluation of college students' psychological status. Different modules taken together can efficiently capture the intricate dynamic properties of psychological states.

4 Experiment

4.1 Dataset

Mostly containing HRV, EDA, and EEG data, the carefully screened and scientifically developed set of multimodal physiological signal data used in this work contains. These data are obtained from publically accessible databases of multimodal mental health research covering the physiological expressions of college students in various psychological states, therefore guaranteeing the variety and representateness of the data.

This work selected data from publicly available multimodal sentiment analysis and mental health research datasets, AMIGOS (a dataset for multimodal research of affect, personality, and mood) and DEAP (a dataset for emotion analysis using physiological signals), so ensuring the dependability and comparability of the data. These datasets cover variations in the psychological state of participants by means of physiological signals gathered under various experimental settings, therefore reflecting changes in the psychological state of viewers watching emotional movies or accomplishing particular activities.

4.2 Data pre-processing

A sequence of standardised data pre-treatment and signal processing methods has been followed to guarantee the correctness and effectiveness of multimodal data:

Low-pass filters help to remove HRV signal high-frequency noise; the wavelet denoising of EDA signals reduces electrode noise interference; Using independent component analysis ICA approach and electromyographic artefact processing including eye electrical interference, remove artefacts from EEG recordings.

Time-domain and frequency-domain aspects of HRV signal feature extraction (such as high-frequency power HF, low-frequency power LF); EDA signal feature extraction: based on signal derivatives, skin conductance response event extraction Calculate the power spectral density of several frequency ranges (like alpha and beta waves) by wavelet transform.

After pre-processing, all modalities' data are standardised so standardising the numerical range and so removing the impact of data scales between several modalities.

4.3 Baseline model

This work performed comparison studies using the following four current models to confirm the efficiency of the approach:

1 Single modal feature extraction model (WTSVM) (Chui et al., 2023): uses CNN to extract features from HRV signals.

2 Early multimodal feature fusion model (ACMNet) (Zhao et al., 2021): directly concatenate multimodal signals and input them into a multi-layer perceptron (MLP) for classification.

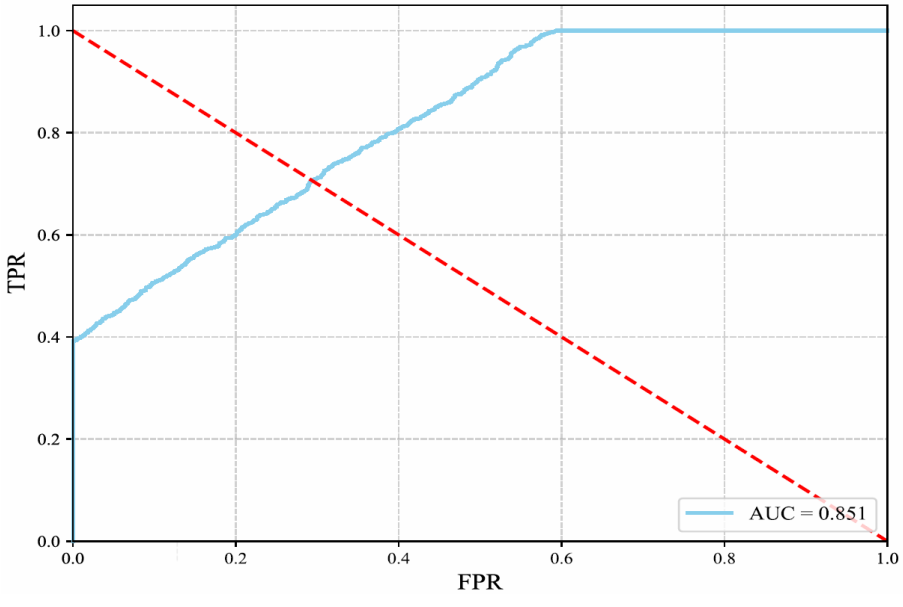
3 Traditional time series prediction model (GRU) (Salem, 2022): using GRU to process time series data of psychological states, it has the ability to handle long-term dependencies.

4 AEVB model (Lopez et al., 2020): making decisions based on a model that conforms to AEVB.

4.4 Comparison of experimental results

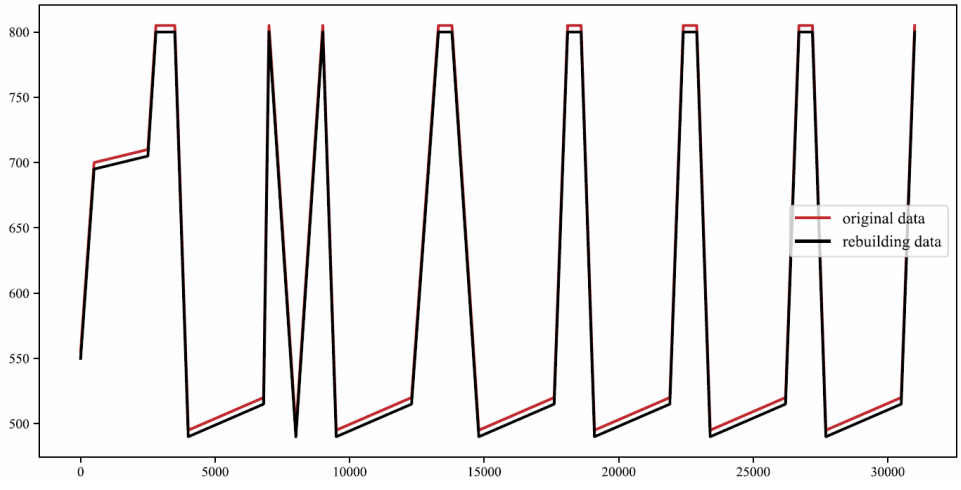
In order to measure the training effect of this model, the experimental scores are plotted in this section, as shown in Figure 3. As shown in the figure, it is the AUC curve in a model experiment, and the AUC value is 0.85, indicating that this chapter has achieved good detection effect. This is because the improved enhanced encoder can obtain better feature extraction effect, and the counter decoder can amplify the reconstruction error.

Figure 3 Experimental score chart



In order to observe the overall reconstruction effect, a sensor is selected in this section to visualise the original data and reconstruction data respectively. The reconstruction effect is shown in Figure 4. The colour of the original data is black, and the colour of the reconstructed data is red. It can be observed that the overall data reconstruction effect is very good, and the complex multi frequency changes of the sensor can also coincide well, which shows the effectiveness of the reconstruction algorithm.

Figure 4 Sensor reconstruction rendering (see online version for colours)



In order to verify the improvement of the training speed of this model, this section conducts a comparative experiment of detection time and training time for the above baseline algorithm, records the detection time of the whole test set and the average time spent on training on each epoch, and draws the comparison of detection time as shown in Table 1 and the comparison of training time as shown in Table 2.

Table 1 Comparison of detection time

Method	*Test duration (seconds)*
WTSVM	3.12
ACMNet	2.7
GRU	2.52
AEVB	1.84
TVLSTM	1.53

Table 2 Comparison of training time

Method	*Training duration (seconds)*
WTSVM	563
ACMNet	105
GRU	22
AEVB	10
TVLSTM	87

5 Conclusions

Combining multimodal physiological signal fusion with deep generation models, this paper suggests a dynamic assessment approach for the psychological condition of college students. Using temporal synchronising and data pre-processing methods, the study

gathered multimodal physiological data including HRV, EDA, and EEG and guaranteed signal consistency and high quality. Based on Transformer structure, the feature extraction network efficiently combines multimodal data, detects the temporal correlation and complementary properties of signals. Enhanced VAE combines LSTM to forecast the dynamic trends of psychological states and develops their latent spatial distribution. This work offers a fresh method for field-based quantitative evaluation in the domain of mental health. Integration of multimodal physiological information with deep learning models not only increases the accuracy of dynamic assessment but also has major consequences for the research of psychological state prediction and intervention mechanisms. To improve the resilience and generality of the model even further, future research will widen application situations and enlarge datasets.

Acknowledgements

This work is supported by the A research project commissioned by the Department of Education of the Inner Mongolia Autonomous Region named: Assessment of College Students' Mental Health Status and Development of Countermeasures across the Region (No. NSZWT202501).

Declarations

All authors declare that they have no conflicts of interest.

References

Aldayel, M., Ykhlef, M. and Al-Nafjan, A. (2020) 'Deep learning for EEG-based preference classification in neuromarketing', *Applied Sciences*, Vol. 10, No. 4, p.1525.

Auerbach, R.P., Alonso, J., Axinn, W.G. et al. (2016) 'Mental disorders among college students in the World Health Organization world mental health surveys', *Psychological Medicine*, Vol. 46, No. 14, pp.2955–2970.

Baltrušaitis, T., Ahuja, C. and Morency, L-P. (2018) 'Multimodal machine learning: a survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp.423–443.

Chen, J., Zhang, Y., Wu, J. et al. (2023) 'SOC estimation for lithium-ion battery using the LSTM-RNN with extended input and constrained output', *Energy*, Vol. 262, p.125375.

Chui, K.T., Gupta, B.B., Torres-Ruiz, M. et al. (2023) 'A convolutional neural network-based feature extraction and weighted twin support vector machine algorithm for context-aware human activity recognition', *Electronics*, Vol. 12, No. 8, p.1915.

Cosoli, G., Poli, A., Scalise, L. et al. (2021) 'Measurement of multimodal physiological signals for stimulation detection by wearable devices', *Measurement*, Vol. 184, p.109966.

Frigant, V. and Jullien, B. (2014) 'Comment la production modulaire transforme l'industrie automobile', *Revue D'économie Industrielle*, Vol. 32, No. 145, pp.11–44.

Goodfellow, I., Pouget-Abadie, J., Mirza, M. et al. (2020) 'Generative adversarial networks', *Communications of the ACM*, Vol. 63, No. 11, pp.139–144.

Greff, K., Srivastava, R.K., Koutník, J. et al. (2016) 'LSTM: a search space odyssey', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 10, pp.2222–2232.

Islam, Z., Abdel-Aty, M., Cai, Q. et al. (2021) 'Crash data augmentation using variational autoencoder', *Accident Analysis & Prevention*, Vol. 151, p.105950.

Kang, H.K., Rhodes, C., Rivers, E. et al. (2021) 'Prevalence of mental health disorders among undergraduate university students in the United States: a review', *Journal of Psychosocial Nursing and Mental Health Services*, Vol. 59, No. 2, pp.17–24.

Khan, S., Naseer, M., Hayat, M. et al. (2022) 'Transformers in vision: a survey', *ACM Computing Surveys (CSUR)*, Vol. 54, No. 10s, pp.1–41.

Kratzert, F., Gauch, M., Klotz, D. et al. (2024) 'HESS opinions: never train an LSTM on a single basin', *Hydrology and Earth System Sciences Discussions*, Vol. 2024, pp.1–19.

Liu, Q. and Liu, H. (2021) 'Criminal psychological emotion recognition based on deep learning and EEG signals', *Neural Computing and Applications*, Vol. 33, No. 1, pp.433–447.

Liu, Y., Sourina, O. and Nguyen, M.K. (2011) 'Real-time EEG-based emotion recognition and its applications', *Transactions on Computational Science XII: Special Issue on Cyberworlds*, Vol. 3, pp.256–277.

Lopez, R., Boyeau, P., Yosef, N. et al. (2020) 'Decision-making with auto-encoding variational Bayes', *Advances in Neural Information Processing Systems*, Vol. 33, pp.5081–5092.

Olivares-Galván, J.C., Georgilakis, P.S. and Ocon-Valdez, R. (2009) 'A review of transformer losses', *Electric Power Components and Systems*, Vol. 37, No. 9, pp.1046–1062.

Posada-Quintero, H.F. and Chon, K.H. (2020) 'Innovations in electrodermal activity data collection and signal processing: a systematic review', *Sensors*, Vol. 20, No. 2, p.479.

Rodríguez-Romo, G., Acebes-Sánchez, J., García-Merino, S. et al. (2022) 'Physical activity and mental health in undergraduate students', *International Journal of Environmental Research and Public Health*, Vol. 20, No. 1, p.195.

Salem, F.M. (2022) 'Gated RNN: the gated recurrent unit (GRU) RNN', *Recurrent Neural Networks: From Simple to Gated Architectures*, Vol. 2, pp.85–100, Springer, Chicago.

Shaffer, F. and Ginsberg, J.P. (2017) 'An overview of heart rate variability metrics and norms', *Frontiers in Public Health*, Vol. 5, p.258.

Sheldon, E., Simmonds-Buckley, M., Bone, C. et al. (2021) 'Prevalence and risk factors for mental health problems in university undergraduate students: a systematic review with meta-analysis', *Journal of Affective Disorders*, Vol. 287, pp.282–292.

Si, C., Yu, W., Zhou, P. et al. (2022) 'Inception transformer', *Advances in Neural Information Processing Systems*, Vol. 35, pp.23495–23509.

Vahdat, A. and Kautz, J. (2020) 'NVAE: a deep hierarchical variational autoencoder', *Advances in Neural Information Processing Systems*, Vol. 33, pp.19667–19679.

Yadav, H. and Thakkar, A. (2024) 'NOA-LSTM: an efficient LSTM cell architecture for time series forecasting', *Expert Systems with Applications*, Vol. 238, p.122333.

Zhao, S., Gong, M., Fu, H. et al. (2021) 'Adaptive context-aware multi-modal network for depth completion', *IEEE Transactions on Image Processing*, Vol. 30, pp.5264–5276.