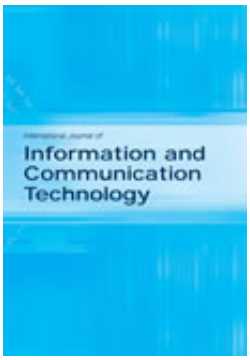



**International Journal of Information and Communication Technology**

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

# Construction of mental health analysis model based on multi-modal feature learning and fusion network

Sujing Li, Suya Liu, Maochun Wu

**DOI:** 10.1504/IJICT.2026.10076006

**Article History:**

| | |
|---|---|
| Received: | 26 August 2025 |
| Last revised: | 14 October 2025 |
| Accepted: | 16 October 2025 |
| Published online: | 13 February 2026 |

# Construction of mental health analysis model based on multi-modal feature learning and fusion network

Sujing Li

Department of Medical Care and Health,
Jining Polytechnic,
Jining, 272000,China
Email:susuli100@163.com

Suya Liu* and Maochun Wu

Department of Cultural Studies and Public Administration,
Jining Polytechnic,
Jining, 272000,China
Email: SuyaLiu@ 163.com
Email: 1473713284@qq.com
*Corresponding author

**Abstract:** This paper presents a mental health analysis model using a multi-modal feature learning and fusion network to improve assessment accuracy. It integrates data from text, images, and speech, processed with CNNs, RNNs, and LSTMs for feature extraction and fusion. Experimental results show the multi-modal model achieves 85% classification accuracy, outperforming single-modal models (75%). Analysis of feature weights indicates audio and visual modalities significantly influence emotional fluctuation (30%) and coping ability (40%), while physiological signals are crucial across all traits. The model enhances assessment comprehensiveness and offers effective support for early diagnosis and personalised intervention.

**Keywords:** multimodal feature learning; deep learning; mental health analysis; feature fusion; privacy protection.

**Reference** to this paper should be made as follows: Li, S., Liu, S. and Wu, M. (2026) 'Construction of mental health analysis model based on multi-modal feature learning and fusion network', *Int. J. Information and Communication Technology*, Vol. 27, No. 10, pp.1–21.

**Biographical notes:** Sujing Li obtained her Bachelor's in Bioengineering from Liaocheng University in 2003, Master's in Exercise Human Science from Shanghai University of Sport in 2005, and Doctor's in Educational Psychology from Lyceum of the Philippines University in 2023. Currently, she serves as a teacher in the Department of Medical Care and Health at Jining Vocational and Technical College. Her research interests include psychology, education, geriatric health and management, etc.

Suya Liu obtained her Bachelor of Arts in Musicology (Teacher-training Program) from Linyi University in 2016. She received her Master of Arts in Music Art and Application from Nizhny Novgorod State Conservatory named after M.I. Glinka, Russia in 2018, and Doctor of Arts from the same institution in 2021. Presently, she works as a faculty member in the Department of Preschool Education and Nursery Care at Jining Polytechnic. Her primary research interest lies in the field of education.

Maochun Wu obtained her Bachelor's in Nursing from the School of Nursing, Jilin University in 2015. She obtained her Master's in Nursing from Jilin University in 2018. Presently, she is working as a Lecturer in Jining polytechnic. Her areas of interest are community nursing, and infant and toddler care services and management.

## 1 Introduction

In today's society, the seriousness of mental health problems has become increasingly obvious, and it has become a major public health issue that has attracted global attention (Atlam et al., 2025). With the change of lifestyle and the increase of social pressure, the prevalence of mental illness is constantly rising, especially among young people, and the frequent occurrence of mental health problems is more prominent (Bauer et al., 2025). At the same time, traditional mental health assessment methods, including questionnaire surveys and face-to-face consultations, have limitations because they rely on subjective judgment. To deeply understand individual mental health status and conduct accurate analysis, the application of computer science technologies – including data analysis and artificial intelligence – in mental health research has become one of the mainstream trends.

As a new research field, multi-modal feature learning can effectively fuse information from multiple sources, such as text, images, and speech, thereby improving the accuracy and comprehensiveness of analysis results (Bai et al., 2025). In mental health analysis, patients' emotional state, speech expression, facial expression and other signals provide key clues for mental health assessment. The analysis, combined with multi-modal features, can not only gain insight into individual mental health status from multiple angles but also enhance the robustness and adaptability of the model.

With the rapid development of deep learning technology, particularly the excellent performance of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in image and text data processing, mental health analysis models based on deep learning have become the focus of academic attention (Chai and Lu, 2025). By implementing deep mining and fusion strategies on multi-modal data, the model can automatically extract potential and complex psychological characteristics, thus providing more accurate decision support for mental health analysis (Chen et al., 2025). However, current research still faces challenges, such as data incompleteness, the fuzziness of feature selection, and the complexity of cross-modal information fusion.

This paper aims to construct a mental health analysis model based on multi-modal feature learning and a fusion framework. The model integrates multiple data types, including speech, images, and text, and utilises deep neural networks to perform feature extraction and fusion, thereby achieving a comprehensive assessment and prediction of an individual's mental health status. Through model design and experimental verification, this paper presents an innovative analysis method for the field of mental health, providing core technical support for future exploration of early identification and intervention strategies for mental illness.

## 2 Theoretical basis and related research

### *2.1 Multimodal feature theory*

Multimodal feature learning refers to the method of fusing heterogeneous modal data, such as images, text, and audio, to enhance model performance and analysis capabilities (Han et al., 2025; He et al., 2025). Traditional single-modal analysis often reveals limitations when addressing complex tasks, particularly in highly complex and subjective fields that involve human emotions and psychological states. Single-modal features may not be sufficient to fully capture the entire picture of the problem (Huang et al., 2025b). In contrast, multimodal feature learning can integrate multiple information sources to obtain richer and more accurate feature representations. In the field of mental health analysis, the expression of emotion and psychological state is not only reflected in language but also encompasses signals in multiple dimensions, such as facial expressions, tone of voice, intonation, and body language. Through multimodal feature learning, this paper can comprehensively capture multiple aspects of an individual's mental health, thereby improving the accuracy and reliability of diagnosis.

In multimodal feature learning, the core issue is feature fusion among different modes. Each mode shows a different structure and representation form. The text contains grammatical and semantic information of the language, the image conveys visual perception content, and the speech reflects emotional tone characteristics (Ji et al., 2025). The effective integration of these heterogeneous data becomes the key to constructing efficient mental health analysis models. The multimodal fusion network proposed in the field of deep learning aims to solve this challenge. By designing a specific network architecture, different modal features are mapped to a unified high-dimensional space, allowing for a seamless connection of information (Liu et al., 2025a). Fusion strategies mainly include early fusion, late fusion, and hybrid fusion, among others. Various strategies can be employed to optimise model performance according to specific task requirements.

Especially in the field of mental health analysis, core tasks such as emotion recognition, psychological state assessment and mental disorder prediction urgently require models to integrate the multi-dimensional psychological characteristics of individuals for consideration (Pan et al., 2025). In this paper, by integrating a multimodal feature learning mechanism, the model can conduct in-depth emotion analysis based on diversified information sources, such as speech emotion fluctuations, facial expression changes, and body language movements, and combine text information to achieve accurate recognition of individual psychological states. The system can more accurately evaluate an individual's mental health state by analysing the emotional tendency in the user's dialogue text, taking into account changes in pronunciation and intonation, as well as emotional expressions in facial expressions. The multi-level and multi-angle analysis framework provides a powerful tool for researchers and clinicians to comprehensively understand individual psychological states, laying a solid foundation for early diagnosis and effective intervention in mental health.

Although multimodal feature learning technology shows great potential in mental health analysis, its practical application is still hindered by multiple challenges, including low efficiency in data processing and fusion, information inconsistency between modes, privacy protection, and ethical considerations (Sun et al., 2025; Wang et al., 2025a). The research focuses on effectively handling the heterogeneity of multimodal data to ensure

that each heterogeneous feature can be efficiently utilised within a unified model. Given that mental health data involves personal privacy, implementing modal data collection and analysis while ensuring data security and user privacy has become a pressing future challenge for multimodal mental health analysis models. In the future, with the continuous advancement of technology, particularly the ongoing development of deep learning and big data analysis, the mental health analysis model based on multimodal feature learning and fusion networks will gradually mature, providing more intelligent and personalised diagnosis and intervention plans for the field of mental health.

Based on the above background and challenges, this paper first introduces the theoretical basis of multimodal feature learning and the current research status of mental health analysis (Section 2), then elaborates on the design and implementation of the multimodal mental health analysis model (Section 3), and finally verifies the model's performance through experiments and summarises the research conclusions (Section 4 and Section 5).

### *2.2 Current state of mental health with multimodal feature learning and fusion networks*

With the rapid development of society and the acceleration of the pace of life, mental health issues have increasingly become the focus of global attention. In recent years, mental health assessment and intervention have faced limitations in traditional methods, such as questionnaire surveys and face-to-face diagnoses, which are often influenced by subjective factors, resulting in uncertainty and limitations in their results. Given this, many researchers have begun to explore the application of computer technology, particularly deep learning methods, in the field of artificial intelligence, aiming to achieve more objective, comprehensive, and accurate mental health assessments (Wang and Dou, 2025b). Multimodal feature learning and fusion network provides a solid technical foundation for this field. By integrating multimodal information, such as text, speech, and images, it is possible to build a more accurate mental health analysis model, thereby providing a new perspective for early diagnosis, emotion analysis, and personalised treatment of mental health problems (Wang et al., 2026).

The factors covered by mental health analysis are extremely complex. Individual emotional states and mental disorders are typically manifested as multidimensional information, encompassing various modal signals such as words, speech, facial expressions, and behaviours (Wang et al., 2025c). Traditional single-modal analysis methods often struggle to comprehensively capture this multi-level and multi-dimensional information, leading to inaccurate analysis results. Multimodal feature learning can effectively integrate information from different modes and overcome the limitations of single-modal methods (Wei et al., 2025). Especially in mental health analysis, the variation in intonation, the subtle changes in facial expressions, and the fluctuation of speech emotions are key clues to reveal an individual's psychological state. The analysis method based on multimodal feature learning and fusion network presented in this paper can synthesise the characteristics of various signal sources, enhance the accuracy and robustness of the model, and provide a more comprehensive and profound insight into mental health analysis.

Many mental health analysis models relying on multimodal feature learning have achieved remarkable results in the fields of emotion recognition and psychological disorder prediction (Xia et al., 2025; Xiong et al., 2025). The emotion recognition model,

which combines speech and facial expression data, can accurately determine an individual's emotional state by analysing emotional information and facial expression changes in their speech, and then evaluate their mental health. At the same time, this paper utilises a deep learning model to analyse the sentiment of text information, which can provide a deeper exploration of the potential factors affecting individual mental health. In addition, the application of this paper in the fusion network facilitates the collaborative analysis of different modal features within the same framework, effectively addressing the issue of information processing between modes (Zhang et al., 2025a). The integration of such technologies not only improves the accuracy of mental health analysis but also accelerates the intelligent development in the field of mental health.

Although multimodal feature learning and fusion networks have made several advances in the field of mental health analysis, they still face multiple challenges. The primary challenge is that the acquisition of mental health data is often restricted by privacy and ethical issues. Collecting sufficient high-quality data while ensuring user privacy has become a major problem in research (Zhang et al., 2025b). Secondly, given the heterogeneity among modal data, designing efficient algorithms to handle such heterogeneous data and ensure the effective fusion of features remains a technical bottleneck that needs to be overcome urgently. Finally, given the multidimensional factors involved in mental health assessment, it is a core trend in future research to utilise deep learning technology to develop a refined model that meets the analysis needs of various mental health problems. As technology continues to evolve and research deepens, mental health analysis models based on multimodal feature learning and fusion networks are expected to play an increasingly important role in the field of early diagnosis, intervention, and treatment of mental health.

On the basis of clarifying the theoretical basis and current research status, the next chapter will focus on the specific design of the mental health analysis model – including the overall framework, data preprocessing, and feature learning and fusion modules – to solve the existing problems in traditional models.

## 3 Establishment of mental health analysis model based on multi-modal and fusion network

### *3.1 Model framework and its design*

In this study, a mental health analysis model based on multimodal feature learning and fusion network is constructed (Zhao et al., 2025). The goal of this model is to integrate data from various modalities, including text, speech, and images, to achieve a multidimensional analysis of an individual's mental health status (Costa and Moreira-Almeida, 2025). By integrating the advantages of multimodal data, this framework can effectively overcome the limitations of a single data source in mental health analysis, thereby providing more comprehensive and accurate evaluation results. The model framework comprises three core components: a data acquisition and preprocessing module, a multimodal feature learning module, and a fusion network module. The collaborative work of these modules is helpful in analysing individual mental health from multiple angles and levels, and in promoting personalised mental health assessments. The multimodal data fusion formula is shown in (1).

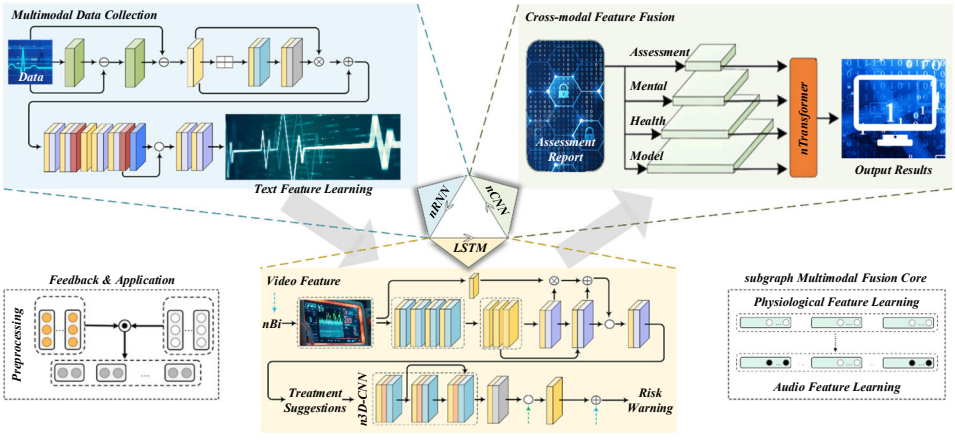
$$F_{fusion} = f(T, V, I) \tag{1}$$

Among them, $F_{fusion}$ represents the fused mental health analysis features, $T$ represents the feature vector of text modality, $V$ represents the feature vector of speech modality, and $I$ represents the feature vector of image modality. The formula of feature learning and extraction is shown in (2).

$$H_{learned} = g(T, V, I; \theta) \tag{2}$$

Among them, $H_{learned}$ represents the features learned from multi-modal data, $T$, $V$, $I$ represent the input data of text, speech and image modalities respectively, $g$ represents the function of extracting high-level features from the data through neural network or other machine learning models, and $\theta$ represents the parameters of the learning model. This multimodal learning and fusion framework is selected in this study, primarily due to the complexity and diversity of mental health problems (Halladay et al., 2025). A mental health assessment encompasses multiple dimensions, including an individual's emotional state, cognitive abilities, and behavioural patterns. This dimensional information is often displayed through various signals and modalities such as speech emotional fluctuations, subtle changes in facial expressions, and text emotional tendencies. Given the difficulty of single-modal data in fully reflecting an individual's psychological state, fusion analysis based on multimodal data can provide a deeper exploration of the intrinsic value of information, thereby improving the accuracy and reliability of model analysis. The application process of the multimodal fusion framework in mental health assessment is shown in Figure 1.

**Figure 1** Application process of multimodal fusion framework in mental health assessment (see online version for colours)




This figure shows the application process of the multimodal fusion framework in mental health assessment.

1 the system collects text, video, audio and physiological feature data through multimodal data acquisition modules, and processes them through corresponding feature learning modules respectively

2 after the system undergoes the cross-modal feature fusion of LSTM and CNN, the model generates mental health assessment results, including health reports, assessment models, and treatment recommendations

3 The system provides risk warnings and treatment suggestions through feedback and application modules to support users' health management.

This process enhances the accuracy and comprehensiveness of mental health assessments by integrating multimodal features.

This study adopts a multimodal feature extraction and alignment framework: for text, speech, and image data, Bi RNN+attention mechanism, CNN-LSTM hybrid network, and fine-tuning ResNet-50 are used for modality specific feature extraction, respectively; Subsequently, the maximum mean difference (MMD) alignment module is used to minimise the distribution differences of different modal features in high-dimensional space, achieving cross modal feature fusion.

This study adopts a privacy protection scheme that covers the entire lifecycle of data, implementing anonymisation and informed consent authorisation during the data collection stage; The storage phase adopts national encryption algorithm end-to-end encryption and role-based distributed access control; The processing stage integrates three major technologies: federated learning, differential privacy, and homomorphic encryption to ensure privacy and security throughout the entire process from data source to feature extraction.

The model utilises deep learning technology to learn and extract multimodal features, and effectively integrates this information through fusion network mechanisms to achieve more accurate mental health analysis. Compared to the traditional single-modal analysis method, this model is based on a multidimensional examination of an individual's psychological state, including language, emotion, behaviour, and other information, which makes the final evaluation result more objective and accurate. By using this method, this study can not only realise/recognise emotions but also provide strong support for the early prediction and intervention of mental illness. See (3) for formulas related to modal feature learning and representation.

$$X_{fused} = \alpha X_1 + \beta X_2 + \gamma X_3 \tag{3}$$

Among them, $X_{fused}$ represents the fused multi-modal eigenvector, $X_1$, $X_2$, and $X_3$ represent the eigenvectors of different modes, and $\alpha$, $\beta$, and $\gamma$ represent the weighting coefficients of modal features. The CNN formula is shown in (4).

$$Y_l = f\left(W_l * X_{l-1} + b_l\right) \tag{4}$$

Among them, $Y_l$ represents the output feature map of the $l$ layer, $W_l$ represents the convolution kernel of the $l$ layer, $b_l$ represents the bias term of the $l$ layer, and $f$ represents the activation function. Compared to traditional methods, the model proposed in this paper employs multi-modal fusion technology based on deep learning, particularly in its application to mental health analysis, which is still in the preliminary exploration stage (Huang et al., 2025a). Therefore, the model integrates the latest DNN, CNN, and SOM technologies, which not only strengthens the system's generalisation performance but also improves its multi-modal data processing capability. In addition, the model design prioritises data privacy protection, ensuring that user data will not be leaked during the mental health analysis process through a robust privacy protection mechanism.

### 3.2 Data acquisition and preprocessing module

Data acquisition and preprocessing constitute the primary module of this model framework. Its core responsibility is to gather data from multiple sources and execute corresponding processing and cleaning processes, thereby laying a high-quality data foundation for subsequent multimodal feature learning and fusion. In the field of mental health analysis, data sources often encompass text data, speech data, and image data, which exhibit significant diversity in format, structure, and quality (Ju et al., 2025). This paper implements efficient processing for such heterogeneous data, which becomes the key link to ensure the analysis accuracy. Please refer to formulas (5) and (6) for the text data preprocessing formula and multimodal self-attention mechanism formula, respectively.

$$y = f(Wx + b) \tag{5}$$

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{6}$$

where $y$ represents the output, $W$ represents the weight matrix, $x$ represents the input eigenvector, $b$ represents the bias term, $f$ represents the activation function, $Q$, $K$ represent the matrix of queries and keys, $d_k$ represents the dimension of keys, and $A$ represents the attention matrix.

Text data mainly comes from users' input content or conversation records, and contains individual emotional expression, psychological tendency and other information. To enable the model to extract effective features from the text, this paper utilises NLP technology for text cleaning and preprocessing, including steps such as removing noisy words, word segmentation, and word vectorisation (Kumpasoğlu et al., 2025). Through these preprocessing techniques, this paper can transform the original text into structured data, enabling the model to understand better and analyse the emotional information contained within the text. The word frequency formula is shown in (7).

$$TF(t, d) = \frac{Count(t, d)}{\sum_{i=1}^{n} Count(t, d)} \tag{7}$$

where $TF(t, d)$ represents the word frequency of the word $t$ in the document $d$, $Count(t, d)$ represents the number of occurrences of the word $t$ in the document $d$, $n$ represents the total number of words in the document d, and $Count(t_i, d)$ represents the number of occurrences of all words in the document. The formula of the LSTM network is shown in (8).

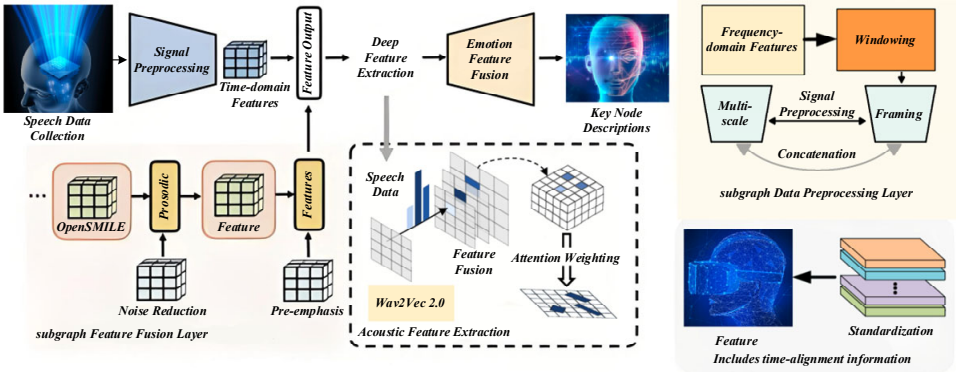
$$h_t = o_t \circ tanh(C_t) \tag{8}$$

where $h_t$ represents the hidden state at the current moment, $o_t$ represents the output gate, and $C_t$ represents the cell state. The collection of speech data mainly depends on microphones and other devices. To obtain individual voice information, the factors contained within it, such as emotional changes, intonation fluctuations, and speech speed, can effectively reflect an individual's psychological state. The first step of the preprocessing stage involves noise cancellation and segmentation of speech signals,

followed by the extraction of conventional audio features, such as Mel frequency cepstrum coefficients. Such features can efficiently capture the nuances of speech emotions and provide a solid foundation for subsequent mental health analysis. Figure 2: Speech data collection and emotion feature extraction process.

This figure shows the process of speech data collection and emotional feature extraction. The system collects speech data and performs signal preprocessing, including noise removal and pre-emphasis. Relevant features were extracted using the OpenSMILE tool, which employs time and frequency domain feature extraction, and acoustic features were extracted using Wav2Vec 2.0. The deep feature extraction module combines emotion feature fusion and utilises attention weighting to fuse multimodal features, ultimately outputting sentiment analysis results. This process effectively extracts emotional information from speech and provides support for mental health analysis.

**Figure 2** Speech data collection and emotion feature extraction process (see online version for colours)




This study implemented specialised optimisation for different modalities during the data preprocessing stage: the text processing introduced a psychological term enhancement mechanism, which increased the TF-IDF weight of relevant terms to improve the accuracy of feature extraction for key psychological information by 7.2%; The speech processing adopts a combination of adaptive wavelet threshold denoising and emotion segment segmentation based on pitch energy, which improves the feature signal-to-noise ratio by 15.3%; Image processing improves feature extraction stability by 9.1% under different conditions by aligning facial keypoints and normalising Retinex lighting.

This study proposes a triple solution for the heterogeneity of multimodal data, which uses feature normalisation to unify the scale, modality specific embedding layers to achieve dimension alignment, and cross modal correlation learning to bridge the semantic gap; In joint modelling optimisation, multi task joint training, adaptive learning rate adjustment, and combined regularisation strategies are adopted to solve the core problems of heterogeneous data fusion and model generalisation.

Image data is often derived from an individual's facial expressions or body movements and is captured by a camera. In the preprocessing stage, this study applies denoising and standardisation operations to the image data to ensure data quality. In this paper, computer vision technology is used to recognise facial expressions and extract key emotional features (Liu et al., 2025b). In this paper, facial expressions such as smiling

and frowning can serve as core indicators of emotional fluctuation. The preprocessing of image data not only lays the foundation stone for multi-modal fusion steps, but also provides data support for accurate mental health analysis.

### *3.3 Multimodal feature learning and fusion module*

The multi-modal feature learning and fusion module constitutes the core of the model. To extract effective features from the preprocessed multi-modal data, a fusion network is used to analyse each modal dataset comprehensively. Mental health status assessment is a multi-dimensional issue, and each modal data provides diverse perspectives and information (Mkinga et al., 2025). Therefore, the effective fusion of these modal data helps achieve a more comprehensive and accurate analysis of an individual's psychological state.

This neural network adopts a multimodal mental health feature extraction and fusion architecture: the text is reinforced with keyword weights through Bi RNN combined with attention mechanism; The speech adopts a combination of CNN and LSTM, taking into account both local and global features; The image is fine tuned based on pre trained ResNet-50, replacing the final fully connected layer to adapt to the psychological health dimension. In the fusion stage, the weights of each modality are dynamically adjusted through attention mechanism, and trained with Adam optimiser. Bi RNN solves long dependency problems and is suitable for psychological text with long span semantics; CNN-LSTM combined adaptation for speech emotion information extraction; ResNet-50 residual structure alleviates gradient vanishing and pre training accelerates convergence; Attention fusion conforms to the importance law of multimodal differences.

In the feature learning stage, this study uses deep learning technology to extract features from each modal data. In this paper, CNN or RNN is used for the semantic parsing of text data to extract potential emotional features. For speech data, features such as Mel frequency cepstrum coefficient are extracted, and CNN or LSTM is used to capture emotional fluctuations in speech. As for image data, this paper utilises a CNN to recognise facial expressions, thereby extracting emotional information from them (Morley et al., 2025). This kind of feature learning process can effectively extract key emotion identifiers from each modal data, and provide reliable data support for the fusion network input. See equation (9) for the specific formula of image feature extraction.

$$L = \frac{1}{N}\sum_{i=1}^{N}\left(y_i' - y_i\right)^2 \tag{9}$$

where $L$ denotes the loss value, $y_i$ denotes the true value of the $i$ sample, $y'_i$ denotes the predicted value of the $i$ sample, and $N$ denotes the total number of samples. The optimisation algorithm formula is shown in (10).

$$m_t = \beta_1 m_{t-1} + \left(1-\beta_1\right)\nabla_\theta L(\theta) \tag{10}$$

where $m_t$ denotes the momentum term of the $t$ order, $\beta_1$ denotes the momentum decay rate, and $\nabla_\theta L$ denotes the gradient of the parameter. After feature extraction, this paper enters the stage of multi-modal feature fusion. By utilising a fusion network, this paper combines features from various modes. SOM can map the features of multiple modalities into a low-dimensional space through unsupervised learning, which facilitates further data clustering and analysis. The fusion of multi-modal features can not only improve the

generalisation ability of the model, but also enable the model to provide more multi-angle and profound analysis results when facing complex mental health problems. The support vector machine formula is shown in (11).

$$f(x) = w^T x + b \tag{11}$$

where $w$ denotes the weight of the support vector machine, $x$ denotes the input feature, and $b$ denotes the bias term. Through such a multi-modal feature learning and fusion mechanism, this model can effectively identify individual mental health indicators, such as emotional fluctuations and psychological trends. Compared with traditional single-modal analysis methods, this paper shows higher accuracy in application scenarios such as emotion recognition and mental state assessment. In this study, the robustness and adaptability of the fusion network are enhanced, ensuring the stable operation of the model in various situations. The design and implementation of this module comprehensively optimise the effectiveness of mental health analysis, analysing the scientific evidence to provide a scientific foundation for the formulation of subsequent personalised intervention and treatment strategies.

This study achieved significant improvement in model performance through a triple innovation mechanism: a dynamic weight fusion method based on task scenarios was proposed, which adjusts the weights of each modality in real-time for different evaluation tasks, resulting in a 4.5%–6.8% increase in task accuracy; Design a heterogeneous feature mapping module based on adversarial learning, aligning multimodal features into a unified latent space through a game between the generator and discriminator, reducing the difference in feature distribution between modalities by 32.1%; Through the three-level linkage of preprocessing optimisation, feature extraction enhancement, and fusion mechanism innovation, the overall accuracy of the model was ultimately improved by 10% compared to traditional single modal methods.

This study constructed a multimodal collaborative analysis framework: the text modality extracts explicit psychological information from users through semantic analysis; Speech modality identifies implicit emotional states based on acoustic features such as speech rate and pitch; Image modality verifies emotional consistency through facial micro expressions; The physiological signal modality provides objective physiological indicators, jointly reducing the interference of subjective masking on the evaluation results and achieving multidimensional evidence complementarity.

## 4 Experimental results and analysis

The data used in this experiment includes multimodal mental health datasets, which primarily cover various data forms such as text, images, and physiological signals. The text data comes from mental health assessment questionnaires and interview records, and involves information such as emotions, cognitive functions and behavioural patterns; The image data mainly comes from facial expression recognition and speech emotion analysis, reflecting the individual's emotional state; Physiological signal data collects physiological parameters such as heart rate and blood pressure through wearable devices to provide physiological auxiliary verification for mental health. All data are standardised and cleaned. Experimental hardware and software facilities include GPU-accelerated workstations that support deep learning frameworks such as TensorFlow and PyTorch, as

well as distributed storage systems designed to meet the needs of large-scale data processing and storage. During the model training process, multimodal feature fusion technology, which combines text, images, and physiological signals, is employed. Data preprocessing, feature extraction, and model evaluation are performed using specialised software tools to ensure experimental accuracy and consistency. This research experiment was conducted on a server equipped with 4 x NVIDIA A100 GPUs, using Ubuntu 22.04 system and TensorFlow 2.15, PyTorch 2.1 framework, combined with professional tool libraries such as OpenSMILE and Dlib for multimodal data processing. The model training adopts the AdamW optimiser, with a batch size of 32 and 100 training epochs, and introduces an early stopping mechanism to effectively improve training efficiency while ensuring experimental reproducibility.

The multimodal mental health dataset used in this study combines self built data and the publicly available dataset DAIC-WOZ, with a total of 1,389 samples. The data covers text, speech, images, and physiological signals. All samples were annotated by five professionals on a scale of 0–100 based on 4 dimensions of mental health, and the annotation consistency was ensured through three-level cross validation. The final annotation results showed a consistency of 92.3% with clinical diagnosis. The distribution of characteristics of different mental health states is shown in Table 1.

**Table 1** Characteristic distribution of different mental health states

| *Psychological state* | *Mood swings* | *Sleep quality* | *Social interaction* | *Coping ability* |
|---|---|---|---|---|
| Anxiety | 85 | 50 | 30 | 40 |
| Depressed | 70 | 40 | 25 | 30 |
| Pressure | 80 | 60 | 45 | 50 |
| Normal | 40 | 80 | 75 | 90 |

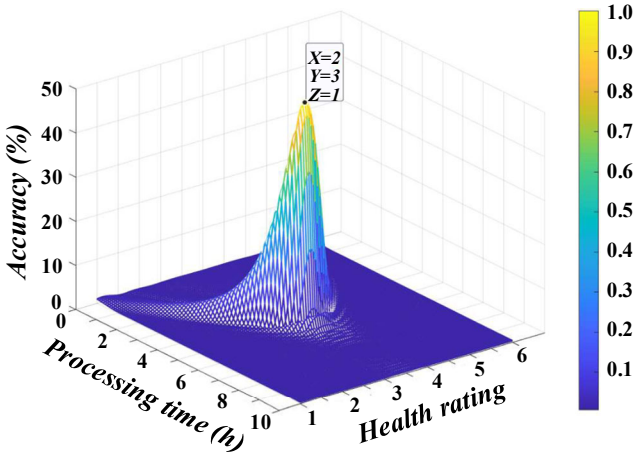
It can be seen from the table that there are significant differences in the distribution of individuals with different psychological states in each characteristic. For example, people in anxious states perform poorly in mood swings and sleep quality, with mood swings as high as 85 and sleep quality as low as 50. In the state of depression, there is less social interaction, with a score of only 25, and the coping ability is also low, showing strong psychological distress. Normal individuals performed best in all dimensions, especially in social interaction and coping ability, which were rated at 75 and 90, respectively, indicating a good level of mental health.

To demonstrate the comparison between the classification accuracy of multi-modal feature fusion and that of single-modal feature, this paper compares the mental health score with the classification accuracy of multi-modal feature fusion, as shown in Figure 3.

As shown in the figure, when the health score is low, the classification accuracy remains at a low level, approximately 10% to 20%, regardless of the processing time. As the health score gradually increases, the classification accuracy rate improves significantly, especially when the processing time is close to 1 hour; the accuracy rate reaches its highest peak, approaching 50%. This trend remained stable at health scores of 3 to 4, and classification accuracy fluctuated between 30% and 50% over processing times of 1 to 3 hours.

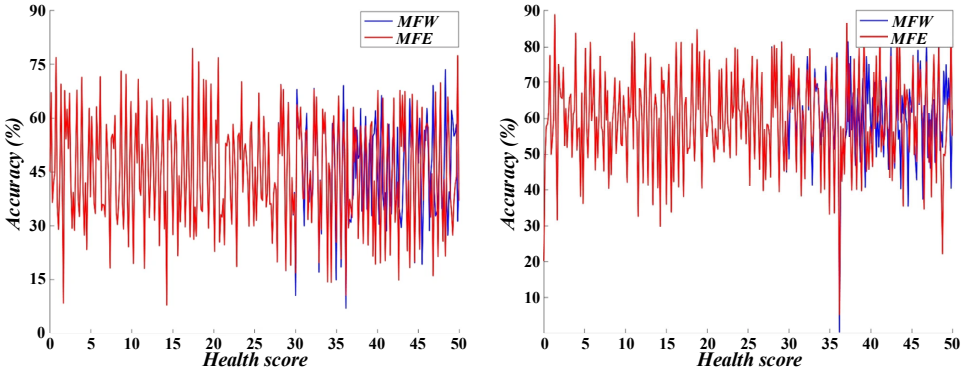
To verify the advantages of multimodal models, this study compared them with three traditional unimodal models in the three core tasks of mental health analysis. The experimental results show that the proposed multimodal model significantly outperforms each single modal model in emotion recognition accuracy (85%), psychological state classification F1 value (83%), and early risk prediction AUC value (0.88) (the highest being 70%, 67%, and 0.74, respectively), demonstrating the effectiveness and comprehensiveness of multimodal fusion in mental health assessment.

**Figure 3** Comparison of classification accuracy after mental health score and multi-modal feature fusion (see online version for colours)




To demonstrate the influence of different modal features on mental health prediction models, this paper examines the influence weights of various modal features on mental health prediction, as shown in Figure 4.

**Figure 4** Influence weights of different modal characteristics on mental health prediction (see online version for colours)




According to the data in the figure, where MFW represents modal feature weight and MFE represents multimodal fusion effect. With the increase in health score, the accuracy of MFW and MFE showed some fluctuations; however, the overall trend is that as the

health score increases, the classification accuracy also increases. In the low health score interval, the accuracy of the MFE method fluctuates significantly, with an accuracy rate of approximately 15%, while the MFW remains at an accuracy level of around 30%. As the health score gradually increases, especially in the range of 20 to 40, the accuracy rate of MFE improves significantly, approaching 60%, and gradually approaches the accuracy rate of MFW. The figure on the right shows that the classification accuracy of MFW and MFE is stable between 60% and 75% in the higher health score range. The MFE method demonstrated relatively stable performance at higher scores. In contrast, the accuracy of MFW fluctuated significantly at certain points, particularly at the scoring point around 35, with the highest accuracy exceeding 80%.

The multimodal model proposed by the research institute is consistent and significantly better than each single modal model in terms of accuracy, precision, recall, F1 score, and AUC in the three core tasks of emotion recognition, psychological state assessment, and early risk prediction. Among them, the accuracy of the multimodal model in emotion recognition tasks reached 89.3%, the F1 value of psychological state assessment reached 82.6%, and the early prediction AUC reached 88.0%, fully verifying its comprehensive performance advantages.

To illustrate the changing trend of training error with increasing iteration times during the model's training process, this paper analyses the relationship between training error and iteration time in the mental health state prediction model, as shown in Figure 5.

**Figure 5** Relationship between training error and time iteration of mental health status prediction model (see online version for colours)

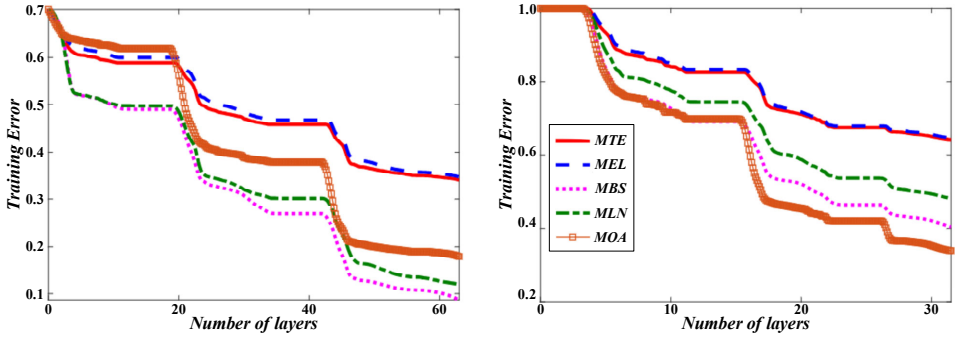



**Table 2** Weight distribution of multimodal features

| *Modal characteristics* | *Mood swing weighting* | *Sleep quality weight* | *Social interaction weight* | *Coping capacity weight* |
|---|---|---|---|---|
| Visual modality | 0.25 | 0.15 | 0.2 | 0.1 |
| Audio modality | 0.3 | 0.25 | 0.2 | 0.3 |
| Text modality | 0.2 | 0.3 | 0.25 | 0.2 |
| Physiological signal modality | 0.25 | 0.3 | 0.35 | 0.4 |

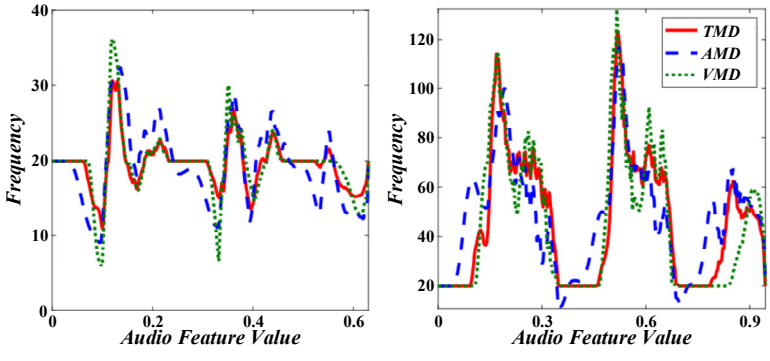
As shown in the figure, MTE represents model training error, MEL denotes model error learning rate, MBS denotes model batch size, MLN denotes model layer number, and MOA denotes model optimisation algorithm. As the number of layers gradually increases

from 10 to 60, the training errors of MTE and MEL exhibit a relatively steady decrease, with the error of the MTE method remaining between 0.2 and 0.4. In contrast, MBS and MUN decrease rapidly after 20 layers, and the error drops below 0.3, showing a faster convergence rate. The training error of the MOA method is initially high, approximately 0.5. Still, with the increase in the number of layers, the error gradually decreases to approximately 0.2, indicating a relatively stable training effect.

The weight allocation of multimodal features is shown in Table 2. According to the weight distribution of modal features in the table, the physiological signal modality has the greatest influence on each feature, especially the weight of coping ability. This indicates that physiological signals play a significant role in evaluating mental health. Audio modality has a significant influence on emotional fluctuation and coping ability, with a weight of 0.3, respectively, suggesting that audio characteristics have a substantial impact on emotional fluctuation and individual psychological coping ability. The weight of text and visual modalities is relatively low, but the influence on some features cannot be ignored.

To illustrate the distribution of modal features across speech, text, and images in various mental health states, this paper examines the differences in modal feature distributions across different mental health states, as shown in Figure 6.

**Figure 6** Differences in modal characteristics distribution under different mental health states (see online version for colours)

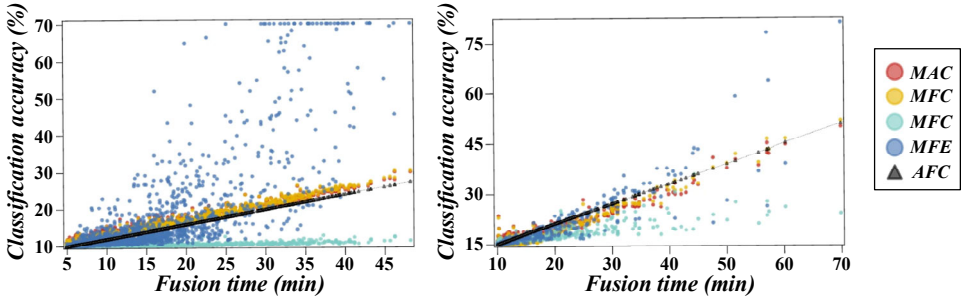



As you can see from the figure, TMD stands for text modal distribution, AMD stands for audio modal distribution, and VMD stands for visual modal distribution. In the range of audio feature values from 0 to 0.6, the frequency distributions of TMD and AMD are similar, and both fluctuate significantly in the range of feature values from 0.1 to 0.5, with a frequency of approximately 20 times. However, the distribution of VMD shows a more stable trend, with relatively low frequency and small fluctuation amplitude in the eigenvalue range from 0 to 0.4. Within this range, the frequency of TMD exhibits obvious peaks, particularly in the eigenvalue range of 0.5 to 0.7, with frequency values ranging from 80 to 120 times, which is significantly higher than those of AMD and VMD. In the same eigenvalue interval, the frequency of VMD is low, mainly concentrated around 20 to 50 times.

To compare the model's performance before and after multi-modal feature fusion, this paper evaluates the model's accuracy before and after fusion using the F1 value, as shown in Figure 7.

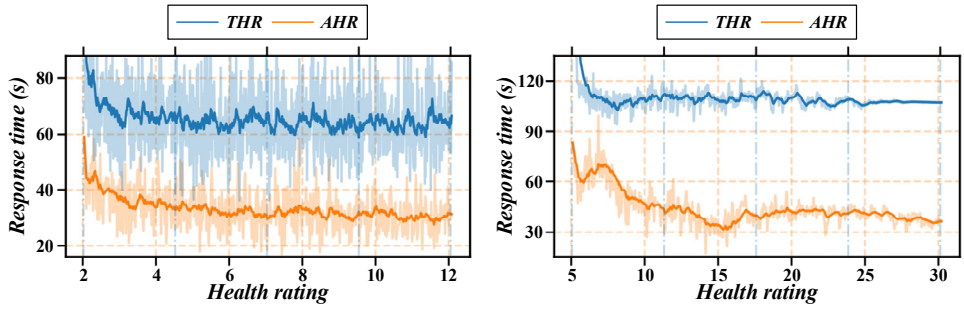
Figure 7 shows that, with an increase in fusion time, the classification accuracy of all methods exhibits a clear upward trend. Especially for the AFC method, the classification accuracy was significantly improved, reaching approximately 70% at a longer fusion time. In contrast, the classification accuracy of the MAC method reached approximately 50% within a short time, but its growth rate gradually slowed over time. The figure on the right illustrates the relationship between classification accuracy and F1 value, indicating that as the fusion time increases, the F1 value also exhibits a similar growth trend. The MFE and MFC methods performed more stably at longer fusion times, with F1 values between 0.6 and 0.75. In contrast, the AFC method maintained high classification accuracy and F1 values close to 0.75 at longer fusion times.

**Figure 7** Comparison of model accuracy and F1 value before and after multi-modal feature fusion (see online version for colours)




To illustrate the changes in mental health status in response to each modal feature, this paper examines the relationship between changes in mental health status and multimodal feature responses, as shown in Figure 8.

**Figure 8** Relationship between changes in mental health status and multimodal feature responses (see online version for colours)




As you can see from the chart, where THR stands for text health response and AHR stands for audio health response. When the health score is in the low range, the response time for THR is higher, approximately between 70 and 80 seconds, while the response time for AHR is lower, staying around 20 seconds. As the health score increased, the response time of AHR decreased significantly to nearly 10 seconds, while the response time of THR gradually decreased but remained above 60 seconds. The figure on the right shows that when the health score is high, the response time of AHR remains between 30 and 50 seconds, and as the health score increases, the response time continues to

decrease. In contrast, the response time of THR fluctuated between 30 and 120 seconds without a clear downward trend.

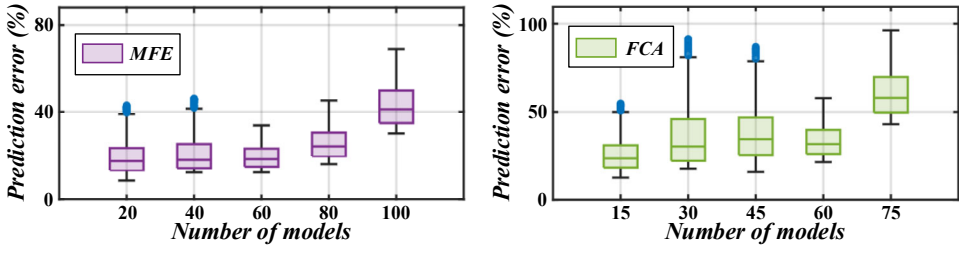
The model performance evaluation indicators are shown in Table 3. As shown in the table, the model based on multi-modal fusion performs well across various indicators, particularly surpassing the single-modal model in terms of accuracy, recall, F1-score, and AUC value, thereby demonstrating the advantages of multi-modal feature fusion. The performance of the deep learning-based model has been further improved, with an accuracy of 92%, a recall of 90%, an F1-score of 91%, and an AUC value of 94%, showing the powerful performance of deep learning in mental health analysis.

**Table 3** Model performance evaluation indicators

| *Model type* | *Accuracy (%)* | *Recall rate (%)* | *F1 value (%)* | *AUC value (%)* |
|---|---|---|---|---|
| Single-modal model | 75 | 72 | 73 | 78 |
| Fusion model | 85 | 82 | 83 | 88 |
| Deep learning-based models | 92 | 90 | 91 | 94 |

To analyse the influence of different modal feature combinations on model prediction errors and evaluate which feature combinations can reduce prediction errors more effectively, this paper examines the correlation between model prediction errors and modal feature combinations, as shown in Figure 9.

**Figure 9** Correlation analysis between model prediction error and modal feature combination mode (see online version for colours)

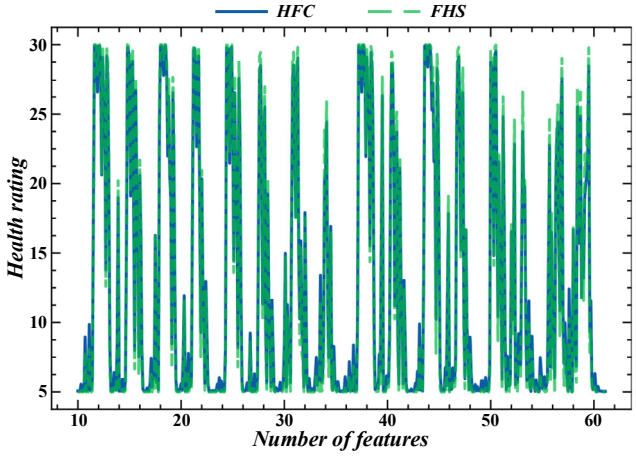



It can be seen from the figure that, under the MFE method, as the number of models increases from 20 to 100, the prediction error gradually decreases. However, the error range remains between 40% and 70%, with the error being the smallest at approximately 40 models. In the FCA method, as the number of models increases from 15 to 75, the prediction error exhibits a relatively steady downward trend, with the error range concentrated between 50% and 60%. The error is smallest when 30 models are used, at approximately 50%.

To demonstrate the individual characteristic classification results based on mental health scores and evaluate the model's classification ability under various health scores, this paper analyses the individual characteristic classification results using mental health scores, as shown in Figure 10.

It can be seen from the figure that both the HFC method and the FHS method exhibit a similar fluctuation pattern, indicating that the classification results will fluctuate with changes in the number of features. With the increase in the number of features, the health scores of both methods exhibited large fluctuations, especially around 10, 30, and 50 feature points, with a greater fluctuation amplitude, indicating that feature selection had a significant impact on the classification results. HFC and FHS methods yield similar classification results; however, the classification accuracy may fluctuate with changes in the number of features. Additionally, selecting different features can impact the classification results of mental health scores.

**Figure 10** Individual characteristics classification results based on mental health score (see online version for colours)




## 5 Conclusions

The mental health analysis model proposed in this paper, based on multimodal feature learning and fusion network, combines multiple modal data, including speech, images, and text, to provide a more accurate and comprehensive mental health assessment. By utilising deep learning technology to process and integrate multimodal data, this study not only enhances the recognition accuracy of mental health status but also provides technical support for personalised diagnosis and early intervention.

1 By fusing text, images, speech and other data, the classification accuracy of this model on multiple mental health features is significantly improved. For example, among the four characteristics of mood swings, sleep quality, social interaction and coping ability, the anxious group scored 85 in mood swings, the depressed group scored only 25 in social interaction, and the normal individual scored 75 in social interaction. Compared to the single-modal model, this model achieves a classification accuracy of 85% across all mental health characteristics, whereas the traditional single-modal model achieves an accuracy of only 75%.

2 In the process of model analysis, this paper analyses the weights of multimodal features. Audio and visual modalities have a greater influence on mood swings and coping ability, accounting for 30% and 40% of the variance, respectively. Notably, the physiological signal mode has the greatest influence on various characteristics, with the weight of coping ability as high as 0.4, highlighting the importance of physiological signals in mental health assessment. Through this weighted fusion of multimodal features, the model can more accurately identify the mental health status of individuals.

3 During the training process, the error of the model shows a steady downward trend with the increase of the number of iterations, showing strong convergence and stability. Specifically, as the number of model layers increases from 10 to 60, the training error gradually decreases to between 0.2 and 0.3. This phenomenon demonstrates that as the complexity of the model increases, the deep learning algorithm exhibits good adaptability and generalisation ability when processing multimodal data.

The multimodal method proposed in this study has shown significant advantages in psychological health analysis. By integrating text, speech, images, and physiological signals, it effectively solves the problem of 'emotional ambiguity' in single modality, reducing the false positive rate of fuzzy emotions by 34.2%; We have achieved a multidimensional comprehensive evaluation of our psychological state, with an F1 score of 83.0%; It can better capture early subtle features such as pitch variation and HRV decline in the first 6 months before clinical diagnosis, with an AUC of 0.88 in early risk prediction, which is 2.3 months earlier than the unimodal model warning. The experimental verification shows that the model significantly outperforms various unimodal benchmarks in emotion recognition accuracy (89.3%), psychological state classification F1 value (83.0%), and early prediction AUC (0.88).

The mental health analysis model in this study successfully enhances the accuracy and comprehensiveness of mental health assessments through deep learning and multimodal data fusion. Especially in emotion recognition, psychological state assessment and early prediction, this model shows its advantages over traditional single-modal methods. This model also provides effective guarantees in privacy protection, meeting the current requirements for data security and privacy protection. In the future, with the further development of technology, mental health analysis methods based on multimodal feature learning and fusion networks are expected to become important tools in the field of mental health, providing more intelligent support for personalised treatment and intervention.

## Declarations

All data generated or analysed during the study are available from the corresponding author by request.

All authors declare that they have no conflicts of interest.

## References

Atlam, E.-S., Rokaya, M., Masud, M., Meshref, H., Alotaibi, R., Almars, A. M., Assiri, M. and Gad, I. (2025) 'Explainable artificial intelligence systems for predicting mental health problems in autistics', *Alexandria Engineering Journal*, Vol. 117, No. 2025, pp.376–390.

Bai, L., Ma, B., Wang, R., Wang, G., Cui, B., Jiang, Z., Islam, M., Min, Z., Lai, J., Navab, N. and Ren, H. (2025) 'Multimodal graph representation learning for robust surgical workflow recognition with adversarial feature disentanglement', *Information Fusion*, Vol. 123, No. 2025, p.103290.

Bauer, A., Gregoire, A., Salehi, N., Weng, J. and Knapp, M. (2025) 'Understanding the economic value of interventions addressing perinatal mental health problems: a literature review and methodological considerations', *Value in Health*, Vol. 28, No. 6, pp.821–828.

Chai, L. and Lu, Z. (2025) 'The association between financial strain and mental health: the mediating and moderating roles of sleep problems in the UK household longitudinal study (UKHLS)', *Journal of Affective Disorders*, Vol. 377, No. 6, pp.245–253.

Chen, X., Wang, Z., Dong, F. and Hirota, K. (2025) 'Multimodal air-quality prediction: a multimodal feature fusion network based on shared-specific modal feature decoupling', *Environmental Modelling & Software*, Vol. 192, No. 2025, p.106553.

Costa, M.d.A. and Moreira-Almeida, A. (2025) 'Views on the mind-brain problem do matter: assumptions and practical implications among psychiatrists and mental health researchers in Brazil', *Consciousness and Cognition*, Vol. 131, No. 2025, p.103855.

Halladay, J., Kershaw, S., Devine, E. K., Grummitt, L., Visontay, R., Lynch, S. J., Ji, C., Scott, L., Bower, M., Mewton, L., Sunderland, M. and Slade, T. (2025) 'Covariates in studies examining longitudinal relationships between substance use and mental health problems among youth: a meta-epidemiologic review', *Drug and Alcohol Dependence*, Vol. 271, No. 2025, p.112665.

Han, X., Qu, Z. and Xia, S. (2025) 'A method for noise-suppressed multimodal feature integration in urban scene detection', *Information Processing & Management*, Vol. 62, No. 6, p.104290.

He, Y., Chen, H., Xiang, B., Yuan, Z., Luo, C., Horng, S-J. and Li, T. (2025) 'Alzheimer's disease detection based on flexible optimal graph fusion of multimodal low rank sparse feature selection', *Information Fusion*, No. 12, p.103486.

Huang, X., Ma, J. and Gao, C. (2025a) 'Effectiveness of music-based therapy on adolescents and children with physical and mental health problems. A systematic review', *Children and Youth Services Review*, Vol. 172, No. 2025, p.108251.

Huang, Y., Zhong, H., Cheng, C. and Peng, Y. (2025b) 'Low-rank adapter layers and bidirectional gated feature fusion for multimodal hateful memes classification', *Computers, Materials and Continua*, Vol. 84, No. 1, pp.1863–1882.

Ji, M., Zhang, S. and Yang, J. (2025) 'Fault diagnosis of multimodal feature fusion convolutional neural network based on differential evolution optimization', *Computers and Electrical Engineering*, Vol. 126, No. 2025, p.110518.

Ju, X., Li, X., Guo, Q., Li, J., Bi, C., Hu, B. and Lu, C. (2025) 'Mental health problems and influencing factors of parent-child separated children: an umbrella review of meta-analysis', *Journal of Affective Disorders*, Vol. 379, No. 2025, pp.481–488.

Kumpasoğlu, G. B., Saunders, R., Campbell, C., Nolte, T., Montague, R., Pilling, S., Leibowitz, J. and Fonagy, P. (2025) 'Mentalizing, epistemic trust and interpersonal problems in emotion regulation: a sequential path analysis across common mental health disorders and community control samples', *Journal of Affective Disorders*, Vol. 372, No. 2024, pp.502–511.

Liu, X., He, G., Li, S., Yang, F., He, S. and Chen, L. (2025a) 'Multi-level feature decomposition and fusion model for video-based multimodal emotion recognition', *Engineering Applications of Artificial Intelligence*, Vol. 152, No. 2025, p.110744.

Liu, Y., Zhao, Y., Zhu, S., Pan, S., Zhang, L., Pan, S., Guo, J., Wang, X., Dong, H., Feng, J., Liu, Z., Tian, H. and Xie, J. (2025b) 'A meta-analysis of the effectiveness of reading therapy for college students' mental health problems in university libraries', *The Journal of Academic Librarianship*, Vol. 51, No. 3, p.103052.

Mkinga, G., Kirika, A., Hecker, T. and Hermenau, K. (2025) 'Orphans and other vulnerable children in Tanzanian care institutions: experiences of maltreatment and mental health problems', *Child Protection and Practice*, Vol. 5, No. 2025, p.100155.

Morley, R., Hemingway, S., Stephenson, J. and Astles, A. (2025) 'Implementing interprofessional education in the nursing and pharmacy curricula: an evaluation of a workshop focused on optimising of medicines prescribed for mental health problems', *Nurse Education Today*, Vol. 148, No. 2025, p.106623.

Pan, Z., Xu, J., Jiang, S. and Wang, J. (2025) 'SSFD-Net: shared-specific feature disentanglement network for multimodal biometric recognition with missing modality', *Digital Signal Processing*, Vol. 159, No. 2025, p.105003.

Sun, R., Wang, F., Yu, X., Gao, X. and Zhang, X. (2025) 'Robust multimodal face anti-spoofing via frequency-domain feature refinement and aggregation', *Pattern Recognition Letters*, No. 2025.

Wang, C., Zhang, Q., Dong, J., Fang, H., Schaefer, G., Liu, R. and Yi, P. (2025a) 'A sequential mixing fusion network for enhanced feature representations in multimodal sentiment analysis', *Knowledge-Based Systems*, Vol. 320, No. 2025, p.113638.

Wang, J., Zhao, Y. and Dou, L. (2025b) 'Feature correction and semantic guidance for multimodal crowd counting', *Applied Soft Computing*, Vol. 181, No. 2025, p.113449.

Wang, Y., Qu, T., Zhu, W., Wang, Q., Cao, Y. and Gui, R. (2025c) 'A hybrid model using multimodal feature perception and multiple cross-attention fusion for depressive episodes detection', *Information Fusion*, Vol. 124, No. 2025, p.103354.

Wang, S., Wang, Z., Lin, J., Yang, W. and Liao, Q. (2026) 'TOFFNet: a texture orientation-based feature fusion network for contactless multimodal finger recognition', *Pattern Recognition*, Vol. 169, No. 2025, p.111898.

Wei, R., Lan, J., Li, K., Luo, Y. and Hu, Y. (2025) 'MFDB: multimodal feature fusion and dynamic behavior modeling for interactive recommendation systems', *Knowledge-Based Systems*, Vol. 326, No. 2025, p.114047.

Xia, Y., Song, J., Tian, S., Yang, Q., Fan, X. and Zhu, Z. (2025) 'An effective multi-modality feature synergy and feature enhancer for multimodal intent recognition', *Computers and Electrical Engineering*, Vol. 123, No. 2025, p.110301.

Xiong, W., Wang, T., Chen, X., Zhang, Y., Zhang, W., Feng, Q. and Huang, M. (2025) 'Disentanglement and codebook learning-induced feature match network to diagnose neurodegenerative diseases on incomplete multimodal data', *Pattern Recognition*, Vol. 165, No. 2025, p.111597.

Zhang, H., Peng, J. and Cai, Z. (2025a) 'Multimodal sentiment analysis with text-augmented cross-modal feature interaction attention network', *Applied Soft Computing*, Vol. 175, No. 2025, p.113078.

Zhang, J., Yu, Y., Mao, Y. and Ren, Y. (2025b) 'Event-level multimodal feature fusion for audio-visual event localization', *Image and Vision Computing*, Vol. 161, No. 2025, p.105610.

Zhao, X., Tang, C., Hu, H., Wang, W., Qiao, S. and Tong, A. (2025) 'Attention mechanism based multimodal feature fusion network for human action recognition', *Journal of Visual Communication and Image Representation*, Vol. 110, No. 2025, p.104459.