Online-Journals.org

iJET International Journal of
Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 24 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i24.45645

PAPER

AI-Assisted Emotion Recognition: Impacts on Mental Health Education and Learning Motivation

Zhiqiang Li(✉)

Department of Mechanical and Electrical Engineering, Hebei Chemical and Pharmaceutical Vocational and Technical College, Shijiazhuang, China

lizhiqiangsdty@163.com

ABSTRACT

With the rapid advancements in artificial intelligence (AI) technology, its deployment in the field of education has gained considerable attention, particularly in the context of mental health education. Addressing the mounting academic and social pressures faced by contemporary students necessitates the utilization of cutting-edge techniques to accurately discern their emotional states and deliver customized learning resources. Existing methodologies for mental health education often fall short due to an over-reliance on educators' experience and observations, as well as challenges in handling complex multimodal data. This research aims to investigate the integration of multimodal audio-visual features using a transformer architecture for emotion recognition. An enhanced probabilistic matrix factorization (PMF) model has been concurrently developed to facilitate tailored content recommendations for students. The goal is to provide a more accurate and effective approach to health education.

KEYWORDS

health education, artificial intelligence (AI), emotion recognition, transformer architecture, multimodal integration, tailored learning content recommendation

1 INTRODUCTION

In the wake of deepening global digitization and the rapid advancement of AI technology, artificial intelligence (AI) has permeated various aspects of daily life [1–10]. Against this backdrop, education, a crucial domain dedicated to nurturing future generations and enhancing the overall quality of society, has naturally emerged as one of the primary applications of AI technology [11–13]. Notably, in the realm of mental health education, AI assistants have shown immense potential. Modern students, whether on physical campuses or within online educational platforms, are faced with unprecedented academic stress, social pressures, and psychological challenges [14, 15]. Under such circumstances, it has become crucial to utilize state-of-the-art technological methods to accurately determine their true emotional states. This will enable us to provide them with timely support and guidance.

Li, Z. (2023). AI-Assisted Emotion Recognition: Impacts on Mental Health Education and Learning Motivation. *International Journal of Emerging Technologies in Learning (iJET)*, 18(24), pp. 34–48. https://doi.org/10.3991/ijet.v18i24.45645

Article submitted 2023-09-12. Revision uploaded 2023-10-25. Final acceptance 2023-10-27.

The fast-paced nature of modern society, combined with the overwhelming amount of information, has resulted in a growing number of students struggling with mental health problems. For educators, the obligation extends beyond monitoring academic progress, encompassing the psychological well-being of their students [16–20]. In this context, integrating AI with mental health education to achieve accurate emotion recognition and personalized learning content recommendations not only presents a technological challenge but also has significant societal implications. Moreover, the integration of AI assistants could introduce a new pedagogical paradigm in the educational sector, potentially enhancing the appeal and effectiveness of the learning experience [21–23].

Traditional mental health education has traditionally relied on the intuition and experience of educators, often neglecting or misinterpreting subtle emotional cues [24, 25]. Concurrently, although existing AI solutions excel in certain applications, the task of deciphering and integrating multimodal data within a complex educational framework remains a significant research challenge. Furthermore, many prevalent recommendation systems tend to favor generic recommendations, disregarding the uniqueness and psychological needs of individual students. This can lead to inappropriate content suggestions.

Two significant research trajectories are explored in this investigation. On one hand, there has been extensive examination of integrating multimodal audio-visual features using the Transformer architecture. The aim is to better capture and interpret students' emotional signals. On the other hand, an improved PMF model has been developed, aiming to protect students' mental well-being while providing personalized learning resources that truly enhance their motivation to learn. The study aims to provide mental health education with a new perspective and methodology. It utilizes AI technology to create a healthier and more effective learning environment for students.

2 MULTIMODAL EMOTION RECOGNITION WITH AI ASSISTANCE

During the learning process, students might experience various emotions, such as anxiety, depression, excitement, or happiness. Through multimodal emotion recognition, it becomes feasible to capture and analyze students' emotional states in real-time, thus potentially identifying psychological health issues at their early stages. Recognizing these emotional states allows educators to provide more personalized emotional support, adapt teaching methodologies or content to better meet the needs of students, improve learning outcomes, and strengthen students' motivation. Trust and connection between educators and students are fostered more effectively when students perceive that their emotional and psychological well-being is a priority. Continuous monitoring and analysis of students' emotional states can proactively predict and prevent potential psychological issues, minimizing their negative impacts on students' learning and lives.

The transformer architecture has demonstrated excellence in various tasks, with its self-attention mechanism facilitating the capture of long-distance dependencies when handling sequential data. To enhance the accuracy of multimodal emotion recognition, this study introduces a model that combines voice and image features using the transformer architecture. By combining multimodal features, the model may achieve higher accuracy in recognizing emotions compared to single-modal approaches. Audio and video data could contain complementary information, enhancing the model's discriminative capabilities.

This model consists of a feature extraction module, a feature fusion module, and an emotion feature classification module (Figure 1). The core functions and operating principles of these three main modules are described in detail below.

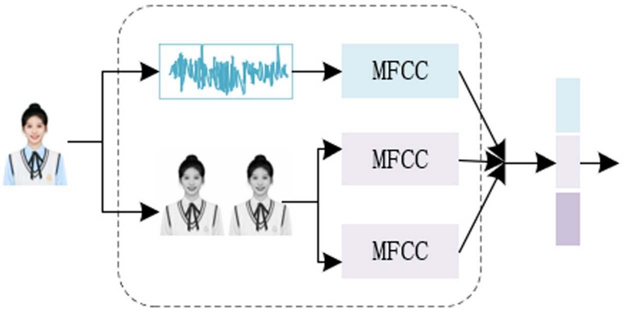

Fig. 1. Architecture of the feature extraction module

Given the vast amounts of raw audio and video data, directly processing them is not only computationally intensive but may also impede the model from acquiring meaningful information. The primary objective of feature extraction is to transform this data into a relatively smaller set of features that represent the essence and meaningful information of the original data. The feature extraction module can capture crucial information for emotion recognition from audio and video. For instance, extracted features from audio might include pitch, rhythm, and intensity, while those from video might focus on facial expressions, eye movements, and other body language. These extracted features serve as input for the feature fusion module, ensuring their structural suitability for subsequent processing and classification.

Audio data first undergoes various preprocessing steps, such as noise reduction and normalization, before extracting essential audio features. These features can capture fundamental properties of sound, such as pitch and rhythm, which are closely associated with emotions. Video data is also subject to preprocessing, which might include adjustments in frame rate and normalization. Key feature extraction from videos typically relies on computer vision techniques. Using these methods, essential emotion-related information, such as eye movement and mouth shape, can be identified and extracted from videos. It is posited that audio features are extracted using the MFCC feature extraction method, represented by d_s, while visual features are derived using the LBP and HOG extraction methods, symbolized by $d_{c,m}$ and $d_{c,g}$, respectively. The number of video frames extracted is represented by l, while the number of feature vectors calculated from audio feature extraction functions and visual feature extraction functions are denoted by "o", "m", and "g", respectively. The original audio-visual samples are represented by T, post-signal processing audio features by T_s, and facial features by $T_{c,m}$ and $T_{c,g}$. This relationship can be expressed as follows:

$$T_s = d_s(ct) = \begin{bmatrix} s_{1,1} & \vdots & s_{1,o} \\ \vdots & \ddots & \vdots \\ s_{l,1} & \cdots & s_{l,o} \end{bmatrix} \tag{1}$$

$$T_{c,m} = d_{c,m}(t) = \begin{bmatrix} c_{1,1} & \vdots & c_{1,m} \\ \vdots & \ddots & \vdots \\ c_{l,1} & \cdots & c_{l,m} \end{bmatrix} \tag{2}$$

$$T_{c,g} = d_{c,g}(t) = \begin{bmatrix} c_{1,1} & \vdots & c_{1,g} \\ \vdots & \ddots & \vdots \\ c_{l,1} & \cdots & c_{l,g} \end{bmatrix} \tag{3}$$

The primary objective of the feature fusion module is to combine features extracted from multiple modalities, such as audio and video, into a unified representation. This integration ensures that information gathered from different modalities is effectively synthesized by the model. A richer and more robust data representation is achieved by combining these diverse features, thereby improving model performance and accuracy. By combining these features, the model generates a cohesive representation that is appropriate for subsequent emotion classification, incorporating all available information. The architecture of the feature fusion module is presented in Figure 2.

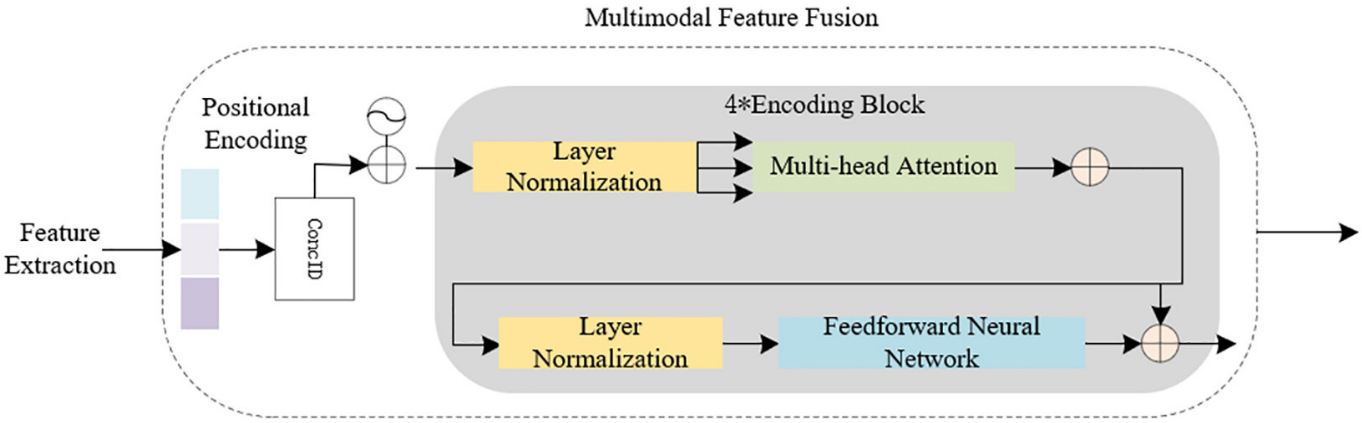

Fig. 2. Feature fusion module architecture

Within this module, feature vectors from both audio and video are concatenated, forming an extended feature vector. This ensures the inclusion of information from both modalities. The combined feature is represented as T_{CA}. The formula for concatenating audio and visual modalities is given by equation (4).

$$T_{CA} = [T_s, T_{c,m}, T_{c,g}] \tag{4}$$

Subsequent to concatenation, these features undergo processing through a convolutional neural network (CNN). CNNs, known for their ability to capture meaningful patterns in local regions, excel at handling time-series data, such as audio features. Further features enter an encoding module where normalization processes occur at each feature layer. This not only stabilizes network training but also accelerates convergence. Utilizing the multi-head attention mechanism of the Transformer, this layer captures intricate relationships among features, particularly long-distance dependencies. Residual connections are introduced at this point to assist the model in effectively training in deeper networks. By adding the outputs of previous layers to subsequent ones, the issue of gradient vanishing is effectively circumvented.

Assuming that class label information is represented by T_{CL}, the input features for the Encoder block by T_0, and linear correlation information by R, with the dimensions of T_{CA}, T_{CL}, and T_{PO} being equivalent, the following relationship is denoted by equation (5). Equations (6) and (7) present the formulas for calculating residual connections and the output of the m-th layer, denoted as T_m.

$$T_0 = \left[T_{CL}; T_{CA}^{1}E; T_{CA}^{2}E; \cdots; T_{CA}^{B}E\right] + T_{PO} \tag{5}$$

The calculation formula for the residual connection is provided in the subsequent equation.

$$T'_m = CC(LN(T_{m-1})) + T_{m-1} \qquad m = 1 \cdots M \tag{6}$$

Assuming the output of the m-th layer is represented by T_m, the calculation formula is as follows:

$$T_m = LMP(LN(T'_m)) + T'_m \qquad m = 1 \cdots M \tag{7}$$

Upon completing these processes, another layer normalization ensures the stability of the network and improves the efficiency of model training. Finally, the processed features are fed into a multi-layer perceptron (MLP) for additional processing and transformation, resulting in the final feature representation used for subsequent emotion classification tasks. If the class label information is T_{CL} and the output post-normalization is t, this relationship is captured by equation (8).

$$t = LN(T_M^0) \tag{8}$$

Emotion state recognition is the main objective of the classification module. Based on the integrated feature representation, the emotional state of students, such as happiness, depression, anxiety, or neutrality, is determined. By categorizing these characteristics, educators and other stakeholders gain valuable insights into the emotional well-being of students. This information can be used to make pertinent educational decisions or interventions. Not only does the classification provide tangible outputs for external users, but it also enhances the transparency of the model's operational process and decision-making logic.

The classification module, as designed in the model, receives feature representations outputted from the feature fusion module. These features, which have undergone prior processing, encapsulate rich emotional information from multiple modalities. Initially, these features are processed through one or more hidden layers. Neurons within these hidden layers utilize activation functions, such as ReLU or Sigmoid, thereby enhancing the model's capability for non-linear representation. This ensures the model's proficiency in discerning complex emotional patterns. Post-processing, the feature information is directed to the output layer. Typically, the output layer comprises multiple neurons, with each neuron corresponding to an emotional category. The deployment of the Softmax activation function ensures that the values outputted by this layer represent the probability distribution of each category. Discrepancies between the model's predictions and the actual emotional labels are evaluated using metrics such as the cross-entropy loss. Subsequently, strategies such as the gradient descent algorithm and backpropagation are employed. In these strategies, the model undergoes self-adjustment based on the loss it experiences, optimizing its weights in the process.

It is posited that the emotion value predicted for the u-th sample is denoted by $t_{-u,j}$. The probability value for the j-th emotional label is denoted by $o_{u,j}$, and the actual emotion value is denoted by $t_{u,j}$. The sample value is symbolized by B, and the total count of emotional categories by J. The model's loss value is determined using the cross-entropy loss function, resulting in the following equation:

$$M = -\frac{1}{B}\sum_{u=0}^{B-1}\sum_{j=0}^{J-1} t_{u,j} \log o_{u,j} \tag{9}$$

Within the probability distribution outputted by the model, the emotional category with the highest probability is selected as the final prediction for the emotional state.

3 PERSONALIZED LEARNING CONTENT RECOMMENDATIONS TO ENHANCE LEARNING MOTIVATION

Intrinsic motivation, which is recognized as a powerful driving force, compels students to actively engage in the learning process. An increase in motivation is often observed when learners find the content intriguing or perceive it as meaningful. Conversely, a mismatch between the learning content and students' interests or needs may result in reduced motivation. Therefore, the significance of providing personalized learning content recommendations aimed at bolstering this motivation cannot be understated. A significant correlation has been identified between learning motivation and psychological well-being. Feelings of frustration and discomfort frequently emerge as primary stressors during the learning process. Personalized recommendations assist learners in identifying optimal learning pathways and resources, thus mitigating unnecessary feelings of defeat and stress. Furthermore, delivering learning content that aligns with students' interests and needs bolsters their satisfaction levels, which in turn promotes psychological health.

The overall framework for the personalized learning content recommendation system, designed to enhance students' learning motivation, can be divided into three distinct phases: student modeling, content modeling, and the recommendation algorithm.

1. Student modeling

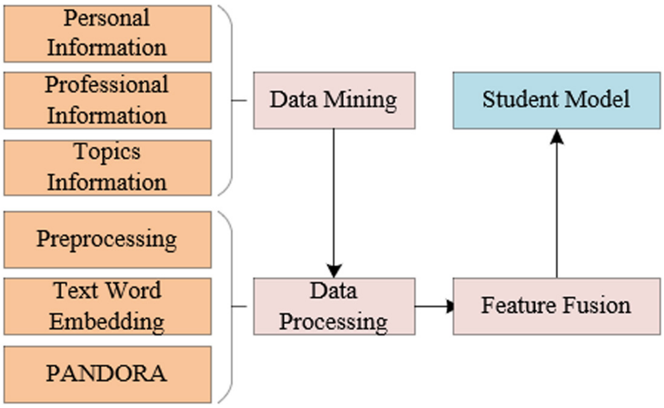

Fig. 3. Workflow of student modelling

For a comprehensive understanding of students' learning habits, motivation, and needs, student modeling is pursued, ensuring more personalized learning content recommendations. Figure 3 illustrates the student modeling workflow. To develop a student model focused on improving learning motivation, essential data about students, including basic information, learning progress, accomplishments, and interaction records, is initially obtained using web scraping techniques. This accumulated data provides an empirical foundation for analyzing student motivation, interests, and needs. Raw data often includes significant noise, missing values, or outliers. In this phase, data is meticulously cleansed, which involves tasks such as removing duplicates, addressing missing values, and identifying and managing outliers. To meet the requirements of the algorithmic model, a series of transformations are performed on the data. These transformations include numeric normalization, vectorization of textual data, and one-hot encoding for categorical data.

Subsequently, after data preprocessing, a meticulous feature engineering process ensues. This involves in-depth analysis of student data to extract key features linked to learning motivation and needs. These features might include variables such as duration of learning, rates of course completion, frequencies of interaction, and test scores. Core features directly related to student motivation and needs, such as learning outcomes and interaction frequencies, are noted. These features reflect the students' learning statuses and intrinsic motivations. Supplementary features could include background information, hobbies, and social networking data. Such attributes further clarify individualized student needs. Through a range of techniques, such as feature selection, feature transformation, and feature fusion, we obtain an integrated feature matrix that combines core and supplementary features. This matrix serves as the foundational input for subsequent algorithms that recommend learning content.

2. Content modeling for learning

The intention behind content modeling for learning is to establish a comprehensive, multi-dimensional representation of learning resources or content. This ensures that recommendation systems can accurately provide learning materials that align with students' needs and interests. Figure 4 delineates the workflow associated with content modeling for learning. Learning content within raw datasets may include course titles, overviews, course content, and related topics. Parsing and cleansing this information is recognized as the initial step in modeling. Topics are often indicative of crucial domains or key points of knowledge within the learning content. Furthermore, manipulating data related to these topics enables more efficient categorization and tagging of the learning content.

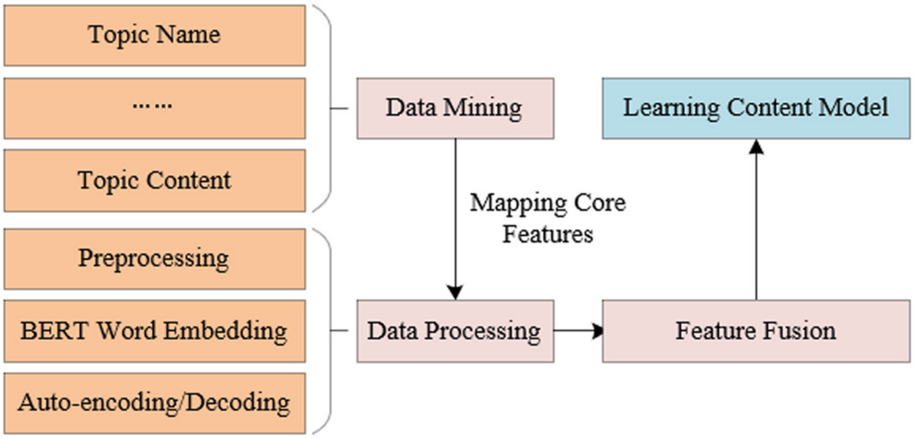

Fig. 4. Workflow of content modelling for learning

Textual feature extraction is subsequently performed on the gathered information. The introduction of word embedding methods based on self-attention mechanisms allows models to assign different weights to each word within the textual data. This approach effectively captures critical information and the contextual relationships within the text. Through autoencoders, a compressed representation of the textual learning content can be obtained. This representation captures the fundamental essence of the text. Decoders, on the other hand, can reconstruct the original text from this compressed representation, thereby verifying the effectiveness of the model. Given the aforementioned processes, key features can be extracted from the learning content text. These features serve as supplementary elements for the learning content and pave the way for subsequent feature fusion.

Lastly, the core feature mapping is complete. Core features may include attributes such as the difficulty level of the learning content, duration, and teaching methodology. It is imperative that these features undergo one-hot encoding in order to be transformed into a format that is suitable for the model. In this phase, the fusion of core features with the previously extracted supplementary features is performed, resulting in a comprehensive feature matrix for learning content. This matrix is intended to serve as the input for the recommendation algorithm.

3. Learning content recommendation algorithm

Within the domain of recommendation systems, probabilistic matrix factorization (PMF) is recognized as a classical matrix decomposition approach. The primary objective is to map users and items onto a shared, low-dimensional space by identifying hidden features, thereby enabling predictions and recommendations. By extending and refining the conventional PMF, the aim of bolstering students' learning motivation through personalized learning content recommendations is achieved. Figure 5 depicts a schematic representation of the framework for the personalized learning content recommendation system.

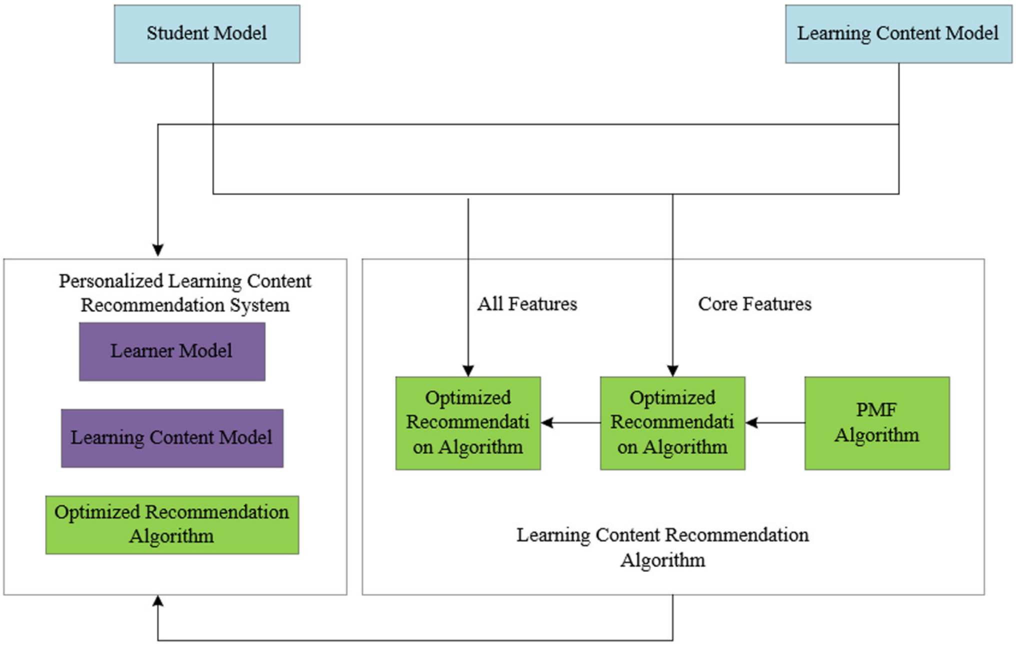
Student Model
Learning Content Model
Personalized Learning Content Recommendation System
Learner Model
Learning Content Model
Optimized Recommendation Algorithm
All Features
Core Features
Optimized Recommendati on Algorithm
Optimized Recommendati on Algorithm
PMF Algorithm
Learning Content Recommendation Algorithm

Fig. 5. Schematic diagram of personalized learning content recommendation system framework

Probabilistic matrix factorization endeavors to identify two low-dimensional matrices, one representing students and the other symbolizing learning content. When multiplied, these matrices can approximate the reconstruction of the original user-item rating matrix. Each row (or column) of these low-dimensional matrices represents the latent feature vector of students or the learning content. Capturing user interests and item attributes is made possible by understanding these underlying features. The result of the proposed algorithm, the probabilistic matrix decomposition, is represented by $R(O(C^u), H(E_u^j))$. The following equation expresses the student model feature matrix.

$$O(\beta I^l, \alpha C_u^l) = O(C_u^l) = \overline{C_u^l} \tag{10}$$

Whereas the learning content model feature matrix is presented as:

$$H(Y^{j\times 1}, E_u^j) = H(E_u^j) = \overline{E_u^j} \tag{11}$$

Assuming a Boolean parameter is denoted by Q_u, Gaussian distribution noise by ω, and the Gaussian function by D, the value becomes true when student l has engaged with topic-k, and false if student l has not engaged with topic-k. Consequently, $R(O(C^u), H(E_u^j))$ aligns with $U(\beta I^l, \alpha C_u^l)$ and $H(Y^{j\times 1}, E_u^j)$ as per the equation.

$$R(O(C_u^l), H(E_u^j), \omega) = \prod_l^b \prod_j^m Q_u \times D(\omega, (R_u, \overline{C_u^l}, \overline{E_u^j})) \tag{12}$$

Multiple dimensions specifically focus on enhancing student learning motivation through psychological health education. These dimensions include emotional and psychological stability, self-efficacy, extended thinking patterns, social connections and a sense of belonging, purpose and goal orientation, psychological need fulfillment, and reflection and self-awareness. Integrating the above dimensions, psychological health education is not only concerned with students' emotional and psychological state. It also explores the development of their confidence, resilience, social skills, and self-awareness—all crucial factors in enhancing student motivation to learn. The recommendation model being constructed must take into account multiple dimensions that enhance students' motivation to learn while also considering the Gaussian noise distribution. This is stated in the following equations:

$$C_u^l = \varepsilon_u + C_z^l \tag{13}$$

$$\varepsilon_u = D(\omega_e Q_u, 0) \tag{14}$$

Further derivations result in:

$$\begin{aligned} R(O(C_u^l), H(E_u^j)) = \omega_e \sum\nolimits_{O_{u,j}} (O(\varepsilon_u + C_u^l)^Y H(E_u^j))^2 + \\ \omega_e O(\varepsilon_u + C_u^l) \sum\nolimits_u O(\varepsilon_u + C_z^l)^{Y^Y} O(\varepsilon_u + C_z^l)^{Y^2} + \\ \omega_e H(E_u^j) \sum\nolimits_k H(E_u^j)^Y H(E_u^j)^2 \end{aligned} \tag{15}$$

4 EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. Performance of the constructed emotion recognition model in unimodal, bimodal, and trimodal emotion recognition

Model	Acc$_7$	Acc$_2$	F1	Corr	MAE
A	42.3	61.5	66.3	0.215	0.824
V	41.5	62.3	64.1	0.248	0.789
T	47.9	78.4	77.5	0.645	0.615
A + V	41.5	61.5	77.5	0.258	0.789
T + V	49.3	77.9	77.1	0.648	0.623
T + V	51.2	78.6	78.9	0.652	0.614
A + V + T	51.3	78.9	81.2	0.669	0.588

From Table 1, results of emotion recognition for unimodal (A, V, T), bimodal (A + V, T + V, T + A), and trimodal (A + V + T) scenarios can be observed. Here, 'A' denotes the audio modality, 'V' represents the video modality, and 'T' signifies the text modality. Within the unimodal scenario, it is evident that the textual modality outperforms others, potentially indicating that textual information offers explicit emotional details in emotion recognition. In bimodal combinations, those that include the textual modality generally exhibit enhanced performance, further emphasizing the importance of the text modality. The trimodal combination of audio, visual, and text (A + V + T) delivers the most optimal results in emotion recognition, suggesting that there may be complementary information among these modalities. Their integration could further enhance the accuracy of emotion recognition. In practical applications, if resources and computational capabilities permit, it is recommended to use the tri-modal combination for optimal performance. If limited by data or computational resources, it is worth considering the use of textual modes or bimodal combinations that include text.

Table 2. Performance comparison between the constructed emotion recognition model and other models

Model	Happy		Sad		Anger		Neutral	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LR	85.6	83.3	81.2	85.3	89.2	83.2	68.9	68.8
SVM	86.2	85.2	77.6	83.4	85.4	84.2	66.9	65.4
MFN	87.6	83.2	82.3	85.2	84.2	84.4	66.8	69.2
NB	87.1	82.3	83.2	83.3	83.3	85.3	67.8	68.2
DNN	87.2	83.1	82.5	86.5	85.1	86.1	65.9	67.5
RNN	85.2	82.1	83.6	87.1	86.3	85.2	69.2	69.3
GRU	86.3	82.6	82.5	75.3	85.2	83.1	68.4	66.4
RAVEN	88.5	86.2	81.2	82.3	85.1	84.2	69.1	68.4
CNN	88.2	83.6	82.8	83.3	83.4	85.5	71.2	71.9
The current model	88.6	88.5	85.6	86.1	85.3	86.2	72.3	72.4

Based on the insights from Table 2, a comparative analysis is conducted to evaluate the performance of the constructed emotion recognition model in comparison to other models. Emotion recognition algorithms, including logistic regression (LR), support vector machine (SVM), matrix factorization (MF), naive bayes (NB), deep neural network (DNN), recurrent neural network (RNN), gated recurrent unit (GRU), RAVEN model, convolutional neural network (CNN), and the research models, have been used in this study. The emotion recognition model developed in this research exhibits outstanding performance across most emotion categories, as indicated by the Acc and F1 metrics. It particularly excels at accurately identifying emotions in the "happy" and "neutral" categories. Compared to other prevalent emotion recognition algorithms, this model demonstrates higher accuracy and stability. This improvement can be attributed to optimized model structure, optimization, feature engineering, or advanced model training strategies. Although the model may not excel in certain metrics, its overall balanced performance is crucial for real-world applications, as data distribution can vary in real-world scenarios. In summary, the emotion recognition model constructed in this research is observed to be a powerful, efficient, and stable tool with significant application value for real-world emotion recognition tasks.

A comparison of performance evaluation metrics for various content recommendation algorithms is drawn from the Table 3. Classic recommendation algorithms such as collaborative filtering (CF), content-based recommendation (CBR), matrix factorization (MF), deep learning recommendation model (DLRM), knowledge graph-based recommendation (KGR), factorization machine (FM), neural collaborative filtering (NCF), adaptive enhanced recommendation, and the proposed algorithm are introduced. Notably, the introduced algorithm excels in MAE, MSE, and RMSE metrics, surpassing all other algorithms, particularly in MSE. While it might not rank the highest in terms of FCP, which is a metric used to evaluate the order of recommended lists, its significant superiority in MAE, MSE, and RMSE, which focus more on prediction accuracy, suggests its clear advantage in terms of accuracy. RLRR also showcases commendable performance in several metrics, particularly in terms of RMSE, indicating its precise recommendation effect. Compared to commonly used recommendation algorithms, the proposed method exhibits distinguished and consistent performance, which can potentially be attributed to its unique model structure, feature engineering, or optimization strategy. Overall, this method offers an efficient, precise, and consistent approach to content recommendation, indicating significant practical value.

Table 3. Performance evaluation metrics for various learning content recommendation algorithms

Algorithm	MAE	MSE	RMSE	FCP
CF	0.612	0.674	0.812	0.624
CBR	0.618	0.682	0.814	0.614
MF	0.589	0.578	0.765	0.628
DLRM	0.589	0.572	0.759	0.615
KGR	0.617	0.612	0.789	0.589
FM	0.623	0.731	0.856	0.614
NCF	0.628	0.726	0.851	0.583
RLRR	0.578	0.569	0.741	0.589
The current algorithm	0.579	0.562	0.752	0.598

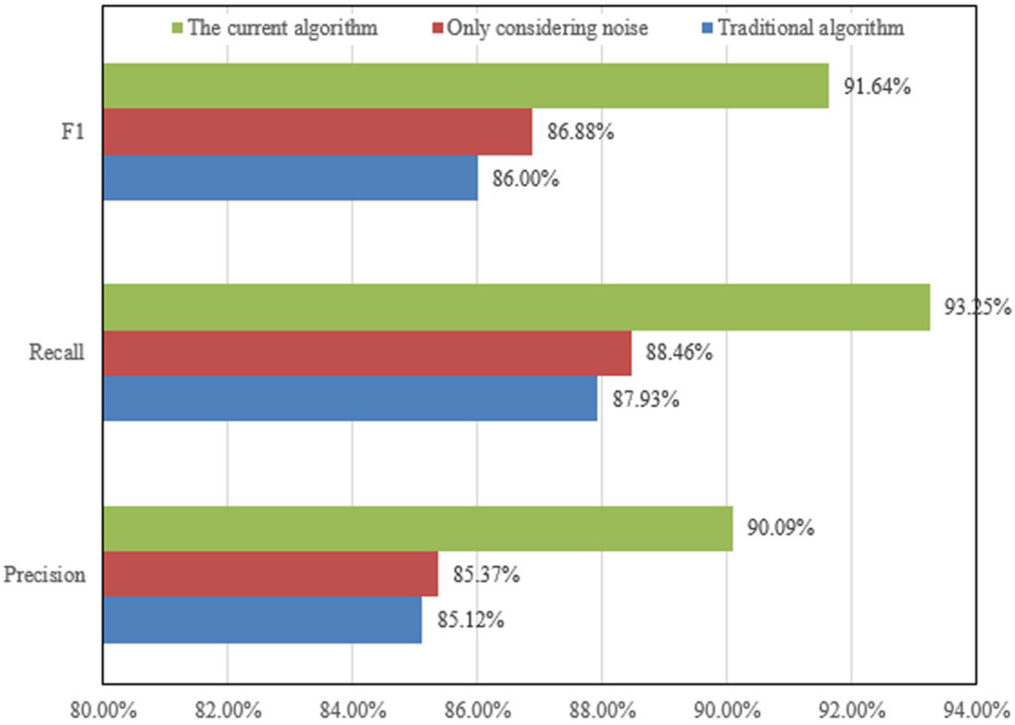

Fig. 6. Comparative analysis of learning content recommendation algorithm before and after improvements

In Figure 6, a comparison of the experimental results of the content recommendation algorithm before and after improvements is depicted. After considering multiple dimensions of student motivation and accounting for Gaussian noise, the proposed algorithm demonstrates significant improvements in all evaluation metrics. This proves the effectiveness and superiority of this comprehensive approach. Methods considering only noise show an improvement over traditional algorithms, yet they fall short when compared to methods that account for multiple dimensions, such as incorporating learning motivation. This suggests that while considering noise is beneficial, taking into account multiple dimensions (e.g., learning motivation) provides greater enhancements. In conclusion, to achieve superior content recommendation results, it is essential to consider not only noise or traditional recommendation factors but also student learning motivations and other relevant aspects. The proposed method in this study presents a significant advantage, providing an effective new approach for content recommendation in learning.

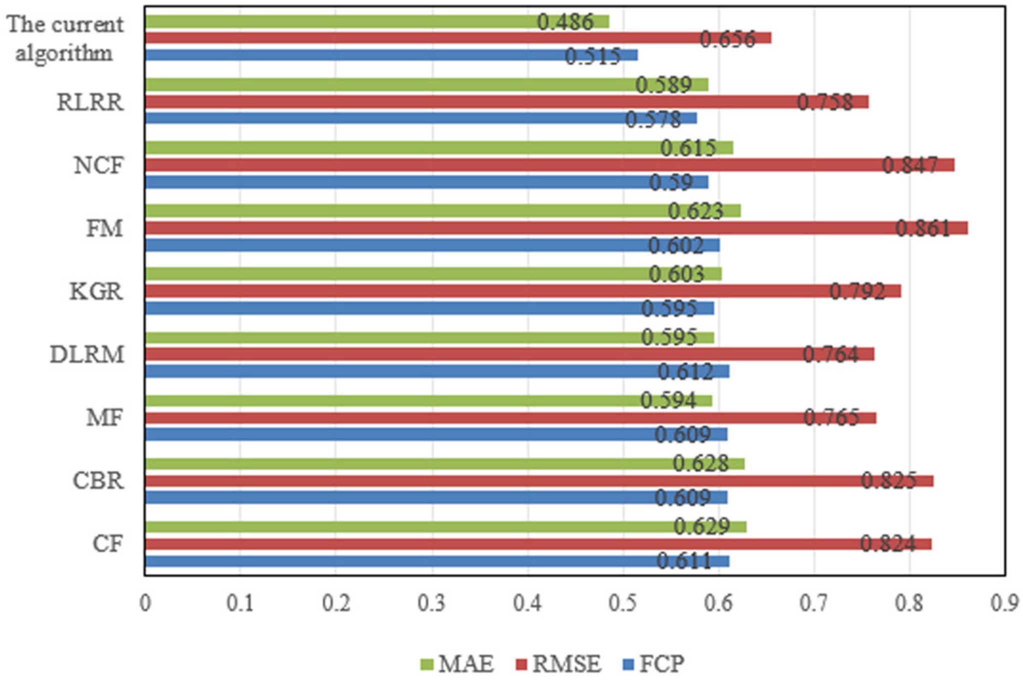

Fig. 7. Bar chart comparison of performance metrics for various learning content recommendation algorithms

Based on Figure 7, an analysis of the performance metrics of various content recommendation algorithms is conducted. The proposed algorithm demonstrates notable excellence in the RMSE and MAE metrics, indicating its superiority in recommendation accuracy. Although its performance on FCP might not be the best, based on the RMSE and MAE results, the proposed algorithm has a significant advantage in terms of accuracy. FCP, on the other hand, focuses more on recommendation orders. Taking all metrics into account, it can be concluded that the proposed algorithm has a significant advantage in terms of recommendation accuracy, making it a powerful tool for providing content recommendations. However, for scenarios that specifically prioritize the order of the recommendation list, additional optimization or the incorporation of other methods may be necessary to improve the FCP score.

5 CONCLUSION

The aim of this research was to explore the application of a multimodal voice-image feature fusion model based on the Transformer architecture in emotion recognition. Additionally, a personalized learning content recommendation system for students was developed. This system is built upon an optimized PMF algorithm model and aims to provide more effective and precise methods for mental health education. Initially, a multimodal emotion recognition model based on the Transformer architecture was introduced. This model comprises three main modules: feature extraction, feature fusion, and emotional feature classification. In a multi-modal context, the study examined the results of emotion recognition using single, dual, and tri-approaches, incorporating audio, video, and text. The investigation particularly emphasized enhancing student motivation. Consequently, a learning content recommendation system was proposed. This system incorporates student modeling, learning object modeling, and the recommendation algorithm. Experimental outcomes indicated that the trimodal combination (A + V + T) achieved the most optimal results in emotion recognition across various metrics. Compared to other prevalent emotion recognition models, the model introduced in this study exhibited superior performance in terms of both accuracy and stability. Among the various recommendation algorithms analyzed for performance, the recommendation algorithm proposed in this research significantly outperformed others in terms of accuracy. Although its performance in the FCP metric might not be optimal, its overall predictive accuracy was found to be significantly superior.

In summary, significant research advancements were achieved in this study in the two pivotal domains of emotion recognition and learning content recommendation. Powerful tools were provided for both fields, establishing the foundation and pointing towards new avenues for future research and applications.

6 REFERENCES

[1] A. S. Ahmad, A. T. Alomaier, D. M. Elmahal, R. F. Abdlfatah, and D. M. Ibrahim, "EduGram: Education development based on hologram technology," *International Journal of Online and Biomedical Engineering*, vol. 17, no. 14, pp. 32–49, 2021. https://doi.org/10.3991/ijoe.v17i14.27371

[2] V. Laciok, A. Bernatik, and M. Lesnak, "Experimental implementation of new technology into the area of teaching occupational safety for industry 4.0," *International Journal of Safety and Security Engineering*, vol. 10, no. 3, pp. 403–407, 2020. https://doi.org/10.18280/ijsse.100313

[3] I. L. F. H. Almutairi, F. L. F. H. Almutairi, and B. F. Alazemi, "Higher education and smart education system: The impact of learning style and environmental characteristics in the state of Kuwait," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 13, pp. 192–199, 2022. https://doi.org/10.3991/ijim.v16i13.30607

[4] S.-H. Ga, H.-J. Cha, and C.-J. Kim, "Adapting internet of things to Arduino-based devices for low-cost remote sensing in school science learning environments," *International Journal of Online and Biomedical Engineering*, vol. 17, no. 02, pp. 4–18, 2021. https://doi.org/10.3991/ijoe.v17i02.20089

[5] R. M. Ramo, A. A. Alshaher, and N. A. Al-Fakhry, "The effect of using artificial intelligence on learning performance in Iraq: The dual factor theory perspective," *Ingénierie des Systèmes d'Information*, vol. 27, no. 2, pp. 255–265, 2022. https://doi.org/10.18280/isi.270209

[6] N. Khamcharoen, T. Kantathanawat, and A. Sukkamart, "Developing student creative problem-solving skills (CPSS) using online digital storytelling: A training course development method," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 11, pp. 17–34, 2022. https://doi.org/10.3991/ijet.v17i11.29931

[7] P. Prommun, T. Kantathanawat, P. Pimdee, and T. Sukkamart, "An integrated design-based learning management model to promote Thai undergraduate computational thinking skills and programming proficiency," *International Journal of Engineering Pedagogy*, vol. 12, no. 1, pp. 75–94, 2022. https://doi.org/10.3991/ijep.v12i1.27603

[8] J. Kwon, "A study on ethical awareness changes and education in artificial intelligence society," *Revue d'Intelligence Artificielle*, vol. 37, no. 2, pp. 341–345, 2023. https://doi.org/10.18280/ria.370212

[9] C. Khentout, K. Harbouche, and M. Djoudi, "Learner to learner fuzzy profiles similarity using a hybrid interaction analysis grid," *Ingénierie des Systèmes d'Information*, vol. 26, no. 4, pp. 375–386, 2021. https://doi.org/10.18280/isi.260405

[10] L. M. Liu, "Analysis on class participation based on artificial intelligence," *Revue d'Intelligence Artificielle*, vol. 34, no. 3, pp. 369–375, 2020. https://doi.org/10.18280/ria.340316

[11] K. Neha, J. Sidiq, and M. Zaman, "Deep neural network model for identification of predictive variables and evaluation of student's academic performance," *Revue d'Intelligence Artificielle*, vol. 35, no. 5, pp. 409–415, 2021. https://doi.org/10.18280/ria.350507

[12] M. Arifin, W. Widowati, and F. Farikhin, "Optimization of hyperparameters in machine learning for enhancing predictions of student academic performance," *Ingénierie des Systèmes d'Information*, vol. 28, no. 3, pp. 575–582, 2023. https://doi.org/10.18280/isi.280305

[13] I. R. Banday, M. Zaman, S. M. K. Quadri, S. A. Fayaz, and M. A. Butt, "Big data in academia: A proposed framework for improving students performance," *Revue d'Intelligence Artificielle*, vol. 36, no. 4, pp. 589–595, 2022. https://doi.org/10.18280/ria.360411

[14] C. Wang, P. Sun, M. Li, and Z. Li, "The application of PREMA model in college mental health education," *Applied Mathematics and Nonlinear Sciences*, 2023. https://doi.org/10.2478/amns.2023.1.00294

[15] B. Khamzina, N. Roza, G. Zhussupbekova, K. Shaizhanova, A. Aten, and B. Aigerim Meirkhanovna, "Determination of cyber security issues and awareness training for university students," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 18, pp. 177–190, 2022. https://doi.org/10.3991/ijet.v17i18.32193

[16] N. T. Van, S. Irum, A. F. Abbas, H. Sikandar, and N. Khan, "Online learning—Two side arguments related to mental health," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 09, pp. 131–143, 2022. https://doi.org/10.3991/ijoe.v18i09.32317

[17] J. Alnuaimi, A. Al-Za'abi, I. A. Yousef, M. Belghali, S. M. Liftawi, Z. F. Shraim, and E. A. M. Tayih, "Effect of a health-based-physical activity intervention on university students' physically active behaviors and perception," *International Journal of Sustainable Development and Planning*, vol. 18, no. 5, pp. 1451–1456, 2023. https://doi.org/10.18280/ijsdp.180515

[18] C. J. Zhu, T. Ding, and X. Min, "Emotion recognition of college students based on audio and video image," *Traitement du Signal*, vol. 39, no. 5, pp. 1475–1481, 2022. https://doi.org/10.18280/ts.390503

[19] S. X. Lu and M. Sim, "Mental health education information analysis system for pre-school students," *Journal of Commercial Biotechnology*, vol. 27, no. 3, pp. 79–89, 2022. https://doi.org/10.5912/jcb1159

[20] Y. Shao and A. J. Abualhamayl, “Differential equation to verify the validity of the model of the whole-person mental health education activity in Universities,” *Applied Mathematics and Nonlinear Sciences*, vol. 7, no. 1, pp. 397–404, 2021. https://doi.org/10.2478/amns.2021.1.00097

[21] X. N. Song and N. L. Lin, “A collaborative education mechanism for college students’ physical and mental health education and safety education under complex background,” *International Journal of Emerging Technologies in Learning*, vol. 17, no. 4, pp. 49–65, 2022. https://doi.org/10.3991/ijet.v17i04.29581

[22] F. Liu and Y. Mai, “Application of computer technology in college students’ mental health education,” in *Proceedings-2022 3rd International Conference on Education, Knowledge and Information Management, ICEKIM*, 2022, pp. 1019–1024. https://doi.org/10.1109/ICEKIM55072.2022.00222

[23] C. Zhang, “Application of K-means clustering algorithm in mental health education evaluation of college students,” in *International Conference on Innovative Computing*, Singapore: Springer Nature Singapore, vol. 935, 2022, pp. 1229–1236. https://doi.org/10.1007/978-981-19-4132-0_168

[24] J. Hong, “Construction of mental health education model for college students based on fine-grained parallel computing programming,” *Scientific Programming*, vol. 2022, p. 4206714, 2022. https://doi.org/10.1155/2022/4206714

[25] Y. Cai and L. Tang, “Correlation analysis between higher education level and college students’ public mental health driven by AI,” *Computational Intelligence and Neuroscience*, vol. 2022, p. 4204500, 2022. https://doi.org/10.1155/2022/4204500

7 AUTHOR

Zhiqiang Li is holds postgraduate degree working as a teacher in the Department of Mechanical and Electrical engineering in Hebei Chemical and Pharmaceutical Vocational and Technical College, China. He graduated from Hebei Normal University. His research interests include psychological crisis intervention of college students’ ideological and political education of college students. He has to his credit two papers published (E-mail: lizhiqiangsdty@163.com; ORCID: https://orcid.org/0009-0000-6050-6108).