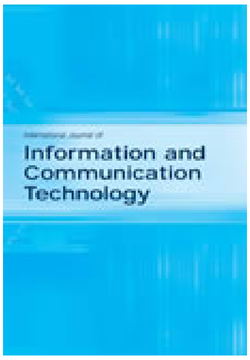


International Journal of
Information and
Communication
Technology

International Journal of Information and Communication Technology

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

Multimodal attentive fusion for emotion recognition model in children's drama

Zhuo Cai

DOI: 10.1504/IJICT.2025.10075260

Article History:

Received:	22 October 2025
Last revised:	15 November 2025
Accepted:	17 November 2025
Published online:	12 January 2026

Multimodal attentive fusion for emotion recognition model in children's drama

Zhuo Cai

School of Music and Dance (SMD),
Changsha Normal University,
Changsha, 410100, China
Email: shuangwaiwai202510@163.com

Abstract: This paper addresses the task of emotion recognition in children's drama performances by proposing an attention-based multimodal feature fusion model. The model extracts fine-grained facial expression features from the visual modality using a pre-trained deep network, and derives Mel-spectrograms and acoustic parameters from the audio modality. These feature streams are then dynamically calibrated and integrated via a cross-modal attention fusion module to capture key emotional cues in dramatic contexts. Evaluated on the public RAVDESS dataset of dramatised speech clips, our model achieves a weighted accuracy of 79.4% and an F1-score of 0.782, demonstrating a significant improvement over feature concatenation-based baseline fusion methods. The results indicate that the model effectively perceives subtle emotional dynamics in theatrical settings, offering a reliable tool for children's affective computing.

Keywords: multimodality; children's theatre; emotion recognition; attentional mechanisms.

Reference to this paper should be made as follows: Cai, Z. (2025) 'Multimodal attentive fusion for emotion recognition model in children's drama', *Int. J. Information and Communication Technology*, Vol. 26, No. 50, pp.113–132.

Biographical notes: Zhuo Cai received his PhD from the Russian National Glinka Conservatory of Music in 2019. He is currently a Lecturer at the Changsha Normal University. His research interests include children's drama, drama education and emotion recognition.

1 Introduction

Affective computing, as an important branch in the field of artificial intelligence, aims to give machines the ability to recognise, understand, interpret and respond to human emotions (Li et al., 2021). It shows great application potential in many fields such as human-computer interaction, intelligent education, and mental health assessment. Especially in the field of children's education, accurate emotion recognition technology can provide key technical support for personalised teaching, emotional intervention for children on the autism spectrum, and immersive interactive entertainment experiences. As a comprehensive art form, children's theatre performance integrates language, vocal tone, facial expression, and body movement, and is one of the most concentrated and rich scenes for children’s emotional expression (Lange and Scheve, 2021). However, children’s emotional expression is often more exaggerated, variable and implicit compared to adults, and their facial expressions and vocal tone changes have higher complexity and uncertainty, which poses a significant challenge to traditional emotion recognition models (Wei et al., 2012).

In recent years, with the rapid development of deep learning technology, unimodal-based emotion recognition research has made great progress. In audio modality, researchers have widely used features such as Mel frequency cepstrum coefficients (MFCCs) and Spectrograms, and utilised convolutional neural networks (CNNs) and recurrent neural networks for temporal modelling, which have achieved considerable results. In terms of visual modality, facial expression recognition techniques based on CNNs have matured and are able to effectively extract emotionally relevant spatial features from still images or video sequences. However, unimodal approaches have inherent limitations: audio information tends to fail in ambient noisy or silent scenes, while visual information is overly sensitive to lighting conditions, occlusion, and head pose. The theatre performance environment is precisely such a variable and complex scene, making it difficult for any single modality to provide a comprehensive and reliable basis for emotion judgment.

In order to overcome the limitations of unimodality, multimodal emotion recognition (MER) emerged and quickly became a mainstream paradigm in current research (Dong et al., 2024). The core idea is to obtain more robust and accurate recognition results by fusing complementary information from different modalities. Pan et al. (2023) proposed a lightweight fully convolutional neural network for efficient extraction of speech emotion features. For the electroencephalogram branch, they proposed a tree-like long and short-term memory model capable of fusing multi-stage features for electroencephalogram emotion feature extraction. Jia et al. (2022) explored the accuracy of MER by using deep learning methods to extract different emotion features from speech, video and motion capture, and designed a matching emotion recognition model – facial motion speech emotion recognition.

Early fusion strategies have mostly focused on feature concatenation or majority voting, which are simple and easy to implement, but tend to ignore the temporal alignment relationship between modalities and the asymmetry of contributions. For example, in theatre performances, the emotion at one moment may be more dependent on exaggerated facial expressions, while another moment may be dominated by changes in pitch. In recent years, fusion strategies based on attentional mechanisms have shown great advantages in that they are able to dynamically assess the importance of different modalities and different time-step features to achieve a more fine-grained fusion. Zhang et al. (2023) proposed a framework for a hybrid audio- and text-based attentional network. The framework combines three different attention mechanisms, such as local intramodal attention, cross-modal attention, and global intermodal attention, which enables effective learning of both intramodal and intermodal emotionally salient features. Li et al. (2024) proposed a sparse interactive attention network (SIA-Net) for MER. In SIA-Net, the sparse interactive attention module mainly consists of intra-modal sparsity and inter-modal sparsity. Intramodal sparsity provides sparse but effective unimodal features for multimodal fusion. Intermodal sparsity adaptively sparsifies intra and intermodal interactions and encodes them as sparse interaction notes. Sparse interaction attention with a small number of non-zero weights then acts on multimodal features to highlight a few but important features and suppress a large number of redundant features.

In addition, intra-modal sparsity and inter-modal sparsity are deeply sparse representations that do not require complex optimisation to make unimodal features and multimodal interactions sparse.

Nevertheless, most of the existing multimodal studies have focused on adult interviews or movie clip scenes, and there has been insufficient research on the specific context of children, especially children's dramatic performances. There are differences between dramatised expressions and natural emotions, and it is often difficult to achieve the desired results by directly applying models designed for adults.

Therefore, this paper is devoted to constructing a MER model based on the attention mechanism applicable to children's dramatic performance scenarios. The main contributions of this paper include

1 designing a dual-stream feature extraction network that extracts discriminative spatial features of facial expressions from video and rich acoustic temporal features from audio, respectively

2 introducing a cross-modal attention fusion module, which is able to adaptively learn the interaction between audio and visual modalities and dynamically weight and integrate the key information, thus effectively capturing the subtle changes in dramatic emotional expressions

3 a systematic experimental validation is conducted on the publicly available Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset of dramatised speech clips.

The results show that compared with the excellent baseline model, the proposed method in this paper achieves significant and reasonable improvement in both recognition accuracy and F1 score, which verifies the effectiveness of the model on this specific task.

2 Relevant technologies

2.1 Attention mechanism foundations and core concepts

Attention mechanism, as an important machine learning method, is inspired by the inherent biological cognitive properties of human beings. In complex information environments, the human brain possesses a unique ability to automatically filter a large amount of redundant information and selectively focus on the most relevant key information for the current task, thus realising efficient allocation of cognitive resources. This cognitive mechanism of selective attention enables humans to maintain efficient information processing in an information overloaded environment (Niu et al., 2021).

From a computational perspective, the attention mechanism is essentially a resource allocation strategy whose core goal is to assign different importance weights to different components when processing information. In deep learning frameworks, this mechanism enables the model to dynamically and selectively focus on specific parts of the input information instead of processing all information equally. This mechanism breaks through the limitation of treating all input features equally in traditional neural network structures and greatly improves the expressiveness and flexibility of the model (Guo et al., 2022).

2.2 Computational principles and implementation of the attention mechanism

The computational process of the attentional mechanism can be understood as a refined information screening process (Soydaner, 2022). The process starts with the evaluation of the relevance of individual elements of the input information to the task at hand. The system generates preliminary attention scores by measuring the strength of association between the query information and each key piece of information through a specific similarity calculation. These scores reflect the importance of different information components for the current processing task.

Subsequently, the system converts these scores into probability distributions through a normalisation process so that the sum of all attention weights is positive and uniform, thus forming formal attention weights (Lu et al., 2023). This conversion process ensures that the weights are standardised and comparable. In the final stage, the system weights and fuses these weights with the corresponding value information to generate a context vector containing the attention information. This context vector is no longer a simple listing of the original information, but an intelligently filtered and enhanced synthesis of the information, which is better suited to the actual needs of the task at hand (Liu et al., 2021).

2.3 Self-attention mechanisms and internal relationship modelling

Self-attention mechanisms are an important evolved form of attention mechanisms, characterised by self-referencing and correlation of information within the system (Choi et al., 2018). In this mechanism, all three elements, query, key and value, are derived from the same input sequence and are generated through different transformations (Brauwers and Frasincar, 2021). This design allows the system to autonomously discover and establish complex correlations between elements within the input sequence.

The core value of the self-attention mechanism lies in its ability to directly capture the dependency between any two positions within the sequence, regardless of the distance of these positions in the sequence. This feature effectively solves the classical problem of distance-dependent decay in long sequence modelling, and provides a new technical path for processing long sequence data (Li et al., 2023). The self-att ention mechanism is not only able to identify local pattern features, but also able to establish global semantic associations, thus realising the deep understanding and characterisation of the input information.

2.4 Multiple attention and multidimensional feature capture

The multiple attention mechanism is an important extension and refinement of the basic attention function (Hernández and Amigó, 2021). The mechanism employs a parallel processing strategy to run multiple independent attention computation processes simultaneously, each focusing on a different aspect and feature dimension of the input information. This parallel architecture is similar to the human cognitive approach of analysing a problem from multiple perspectives at the same time (Li et al., 2020a).

Each attention head is responsible for extracting information from a particular representational subspace, focusing on different feature dimensions of the input data. These attention heads work independently to produce their own information filtering results (Li et al., 2020b). Eventually, the system integrates and fuses these decentralised attention results to form a comprehensive output (DeRose et al., 2020). This design enables the model to capture multiple types of associations and features in the input information simultaneously, greatly enhancing the expressiveness of the model and the comprehensiveness of feature capture.

2.5 Theoretical value and significance of attentional mechanisms

The proposal and development of the attention mechanism has had a profound impact on the field of deep learning (Ghaffarian et al., 2021). From the theoretical level, this mechanism provides an explicit information selection mechanism for neural network models, enabling the models to mimic the human cognitive attention allocation process. This mechanism not only improves the performance of the model when dealing with long sequences and complex data, but more importantly enhances the interpretability of the model (Rodriguez et al., 2019).

By analysing the distribution pattern of the attention weights, researchers can visualise the information regions and features that the model focuses on during the decision-making process, which provides a valuable window into understanding the internal working mechanism of complex neural networks (Lv et al., 2022). The successful application of the attention mechanism promotes the evolution of deep learning models from the black-box style to the direction of explainability and comprehensibility, and lays an important theoretical foundation for the construction of more transparent and reliable artificial intelligence systems.

3 Affect recognition model based on cross-modal attention fusion

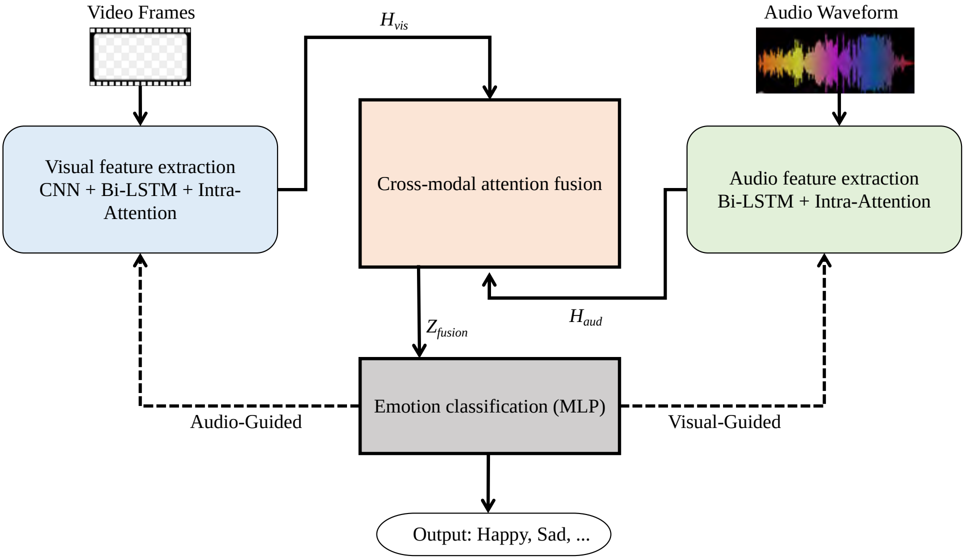
The core architecture of the multimodal feature fusion-based emotion recognition model for children's drama proposed in this paper aims to efficiently process and fuse the temporal information from visual and audio modalities, and ultimately realise the accurate classification of rich emotions in children's drama performances. The overall framework of the model consists of four core components: visual feature extraction module, audio feature extraction module, cross-modal attention fusion module, and emotion classification module. The overall design of the model follows the logical process from bottom-up, from feature extraction to advanced semantic fusion, and its overall structure is shown in Figure 1.

3.1 Visual feature extraction module

Children's emotions are mainly conveyed through rich and exaggerated facial expressions in theatre performances. In order to capture these subtle and dynamically changing visual information, this model employs a visual feature extraction process based on deep CNNs (Kheradpisheh et al., 2018). The process first pre-processes the input raw video sequence for face detection and alignment, and subsequently extracts highly discriminative facial representations from each frame.

Given a video clip containing T frames $V = \{v_1, v_2, \dots, v_T\}$ we first process each frame using an advanced face detector such as MTCNN (Ku and Dong, 2020) v_t to accurately localise and crop out the face region images f_t This pre-processing step effectively removes the background interference, allowing the model to focus on the facial information that is most relevant to emotion.

Figure 1 Methodology structure diagram (see online version for colours)


Video Frames
H_{vis}
Audio Waveform
Visual feature extraction CNN + Bi-LSTM + Intra-Attention
Cross-modal attention fusion
Audio feature extraction Bi-LSTM + Intra-Attention
Z_{fusion}
H_{aud}
Audio-Guided
Emotion classification (MLP)
Visual-Guided
Output: Happy, Sad, ...

Subsequently, we input the cropped face image sequence into a deep convolutional neural network pre-trained on a large face dataset. After removing its top fully connected classification layer, the network acts as a powerful feature extractor. For each frame of a face image f_t, the network outputs a high-dimensional deep feature vector h_t^{vis} whose dimension is denoted as D_v:

$$h_t^{vis} = CNN\left(f_t; \Theta_{CNN}\right), t = 1, 2, \ldots, T \tag{1}$$

where Θ_{CNN} denotes the parameters of the pre-trained *CNN* network. Up to this point, we have obtained a sequence of visual frame-level features $H^{vis} = \left\{h_1^{vis}, h_2^{vis}, \ldots, h_T^{vis}\right\}$.

However, the expression of emotion is a continuous and dynamic process, and a single frame image cannot capture its contextual information evolving over time. For this reason, we introduce a bi-directional long short-term memory (Bi-LSTM) network to model the temporal dependence of visual features (Suebsombut et al., 2021). The Bi-LSTM is able to capture the contextual information of the sequence in both forward and backward directions simultaneously, which is crucial for understanding the onset, peak, and fade processes of expressions.

Visual temporal feature modelling

$$\overrightarrow{h_t^{vis\_lstm}} = \overrightarrow{LSTM}\left(h_t^{vis}, \overrightarrow{h_{t-1}^{vis\_lstm}}; \Theta_{\overrightarrow{LSTM}}\right) \tag{2}$$

$$\overleftarrow{h_t^{vis\_lstm}} = \overleftarrow{LSTM}\left(h_t^{vis}, \overleftarrow{h_{t-1}^{vis\_lstm}}; \Theta_{\overleftarrow{LSTM}}\right) \tag{3}$$

Visual temporal feature splicing:

$$h_t^{vis\_bi} = \left[\overrightarrow{h_t^{vis\_lstm}}; \overleftarrow{h_t^{vis\_lstm}}\right] \tag{4}$$

where $\overrightarrow{h_t^{vis\_lstm}}$ and $\overleftarrow{h_t^{vis\_lstm}}$ represent the hidden states of the forward and backward LSTM at moment t, respectively, and [;] denotes the vector splicing operation. Finally, we obtain the temporal context-rich visual feature sequence $H^{vis\_bi} = \left\{h_1^{vis\_lstm}, h_2^{vis\_lstm}, \ldots, h_T^{vis\_lstm}\right\}$ with dimension D_{vL}.

3.2 Audio feature extraction module

Parallel to the visual modality, the audio channel provides indispensable paralinguistic information such as pitch, volume and speech rate, which are all important cues for emotion recognition.

First, the raw audio signal synchronised with the video is pre-processed, including standard steps such as pre-emphasis, frame-splitting, and windowing. Subsequently, we extract two complementary acoustic features: the MFCCs and the log-mel spectrogram (Nguyen et al., 2023). The MFCCs are able to characterise the spectral envelope of the sound well, which is closely related to human auditory perception; while the Log-Mel Spectrogram retains richer time-frequency information.

Let the audio signal be divided into T time windows (aligned with the video frames), and each time window is extracted to obtain an MFCC feature vector a_t^{mfcc} and a Log-Mel feature vector a_t^{mel} . We splice them to form a joint audio feature vector a_t at time t. The MFCC feature vector a_t^{mfcc} and the Log-Mel feature vector a_t^{mel} are extracted from each time window:

$$a_t = \left[a_t^{mfcc}; a_t^{mel}\right] \tag{5}$$

As a result, we obtain the original audio feature sequence $A = \{a_1, a_2, \ldots, a_T\}$ with dimension D_a.

Similar to visual features, audio emotion information has a strong temporal dependency. Therefore, we similarly employ a bi-directional LSTM network to learn the temporal context model of audio features (Graves and Schmidhuber, 2005).

$$\overrightarrow{h_t^{aud\_lstm}} = \overrightarrow{LSTM}\left(h_t, \overrightarrow{h_{t-1}^{aud\_lstm}}; \Theta_{\overrightarrow{LSTM_a}}\right) \tag{6}$$

$$\overleftarrow{h_t^{aud\_lstm}} = \overleftarrow{LSTM}\left(h_t, \overleftarrow{h_{t+1}^{aud\_lstm}}; \Theta_{\overleftarrow{LSTM_a}}\right) \tag{7}$$

$$h_t^{aud\_bi} = \left[\overrightarrow{h_t^{aud\_lstm}}; \overleftarrow{h_t^{aud\_lstm}}\right] \tag{8}$$

Finally, we obtain the audio feature sequence $H^{aud\_bi} = \left\{h_1^{aud\_bi}, h_2^{aud\_bi}, \ldots, h_T^{aud\_bi}\right\}$ rich in temporal context with dimension D_{aL}.

3.3 Cross-modal attention fusion module

This is the core innovation of the model in this paper. Simple feature splicing or weighted averaging cannot dynamically capture the differences in the contributions of different modalities to the affective state at different moments. This module introduces a cross-modal attention mechanism, which aims to allow the features of one modality to dynamically guide the feature selection and fusion of another modality (Chen et al., 2022).

The module is divided into two phases: intra-modal attention (IMA) and inter-modal attention.

First, we compute self-attention for visual and audio feature sequences separately. Taking the visual modality as an example, we compute an attention weight vector that measures the importance of each time-step feature to the final visual representation.

Attention weights within visual modality.

$$e_t^{vis} = u_v^T \cdot tanh\left(W_v \cdot h_t^{vis\_bi} + b_v\right) \tag{9}$$

Normalisation of attentional weights within the visual modality:

$$\alpha_t^{vis} = \frac{exp\left(e_t^{vis}\right)}{\sum_{j=1}^{T} exp\left(e_j^{vis}\right)} \tag{10}$$

where W_v is a weight matrix, u_v is a context vector, and b_v is the bias term, all of which are trainable parameters α_t^{vis}. That is, the attentional weights of the visual features at moment t. The weights of the visual features are summed to obtain a visual context vector. The weighted summation yields the visual context vector c^{vis} weighted by the intramodal attention:

$$c^{vis} = \sum_{t=1}^{T} \alpha_t^{vis} h_t^{vis\_bi} \tag{11}$$

Similarly, we perform exactly the same operation on the audio feature sequence H^{aud_bi} to obtain the weighted audio context vector c^{aud}.

Next, we perform inter-modal attention interaction. We design bidirectional cross-modal attention (Rabinovich et al., 2013).

Audio as query, visual as key: we use the context vector c^{aud} of audio as a query, which is computed with the features $h_t^{vis\_bi}$ of each time step in the visual sequence, to generate a set of audio information-guided visual attention weights.

Audio-guided visual attention scores:

$$s_t^{a->v} = v_{a->v^T} \cdot tanh\left(W^q_{a->v^{caud}} + W^k_{a->v} h_t^{vis\_bi} + b_{a->v}\right) \tag{12}$$

Audio-directed visual attention weighting.

$$\beta_t^{a->v} = \frac{exp\left(s_t^{a->v}\right)}{\sum_{j=1}^{T} exp\left(s_j^{a->v}\right)} \tag{13}$$

Audio-guided visual representations:

$$z^{a->v} = \sum_{t=1}^{T} \beta_t^{a->v} h_t^{vis\_bi} \tag{14}$$

Vector $z^{a \to v}$ represents the most relevant visual information to the current audio sentiment, guided by the audio information and filtered from the visual modality.

Visual as query, audio as key value: similarly, we use the visual context vector as a query to guide the filtering of audio features.

Visual-guided audio attention score:

$$s_t^{v->a} = v_{v->a}{}^{T} \cdot tanh\left(W_{v->a}^{q} c^{vis} + W_{v->a}^{k} h_t^{aud\_bi} + b_{v->a}\right) \tag{15}$$

Visually guided audio attention weighting:

$$\beta_t^{v->a} = \frac{exp\left(s_t^{v->a}\right)}{\sum_{j=1}^{T} exp\left(s_j^{v->a}\right)} \tag{16}$$

Visually guided audio representation:

$$z^{v->a} = \sum_{t=1}^{T} \beta_t^{v->a} h_t^{aud\_bi} \tag{17}$$

Vector $z^{v \to a}$ represents the audio information most relevant to the current visual emotion filtered from the audio modality guided by the visual information.

Finally, we splice the two original intramodal context vectors c^{vis}, c^{aud} and two new representations $z^{v \to a}$, $z^{a \to v}$ after cross-modal guidance to form the final multimodal fusion feature vector z^{fusion}:

$$z^{fusion} = \left[c^{vis}; c^{aud}; z^{a->v}; z^{v->a}\right] \tag{18}$$

This vector contains both pure modal information, significant information within modalities, and complementary information guided by intermodal interactions, forming the final basis for sentiment classification.

3.4 Sentiment classification module

The fused high-dimensional feature vectors z^{fusion} are fed into a simple multilayer perceptron classifier for sentiment classification. This classifier usually consists of one or more fully connected layers and uses dropout techniques to prevent overfitting (Castro et al., 2017).

Sentiment classification:

$$y_{pred} = Softmax\left(W_y \cdot Dropout\left(ReLU\left(W_f \cdot z^{fusion} + b_f\right)\right) + b_y\right) \tag{19}$$

where W_f, b_f, W_y, b_y is the trainable weight and bias of the classifier. y_{pred} is the probability distribution of sentiment categories predicted by the model.

The training objective of the model is to minimise the cross-entropy loss function between the predicted probability distribution and the true label.

The loss function is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}^{true} log\left(y_{i,c}^{pred}\right) \tag{20}$$

where *N* is the number of training samples, *C* is the number of sentiment categories, $y_{i,c}^{true}$ is the true label of sample *i* on category *c*, and is the probability that sample *i* belongs to category c as predicted by the model.

In summary, this chapter details the MER model proposed in this paper. The model extracts in-depth temporal features of audiovisual modalities through a well-designed two-stream network, and innovatively introduces a two-way cross-modal attention fusion mechanism to dynamically and adaptively capture complex inter-modal interactions, which lays a solid foundation for the final accurate emotion recognition.

4 Experimental results and analyses

In order to comprehensively assess the effectiveness and superiority of the MER model based on cross-modal attention fusion proposed in this paper, we designed and conducted a series of rigorous experiments. This chapter will elaborate on the datasets used in the experiments, the specific data preprocessing process, the performance evaluation metrics employed, the baseline models used for comparison, and the detailed experimental results and analysis. Through in-depth discussion of the results and ablation experiments, we systematically validate the contributions of each component of the model.

4.1 Experimental setup

A high-quality dataset is the basis for effective model training and fair performance comparison. In this study, RAVDESS, a widely used and well-recognised public dataset in the field of affective computing, is selected. The dataset contains audio-visual recordings of 24 professional actors performing eight different emotions, each with two linguistic intensities, with very high performance quality and emotional purity. Although it was not specifically designed for children, the exaggerated and dramatic expressions used by the actors in performing the ‘strong’ emotions are highly consistent with the emotional outwardness of children’s theatrical performances, and thus are very suitable for use as an experimental baseline for this study. From the ‘strong’ intensity clips of all the actors, we selected ‘happy’, ‘sad’, ‘angry’ ‘sadness’, ‘happiness’, ‘anger’, ‘fear’, ‘surprise’, and ‘neutrality’, which are the six most common emotion categories in children’s theatre, with a total of 864 valid samples. We performed a hierarchical division according to actor IDs to ensure that different samples of the same actor would not appear in the training and testing sets at the same time, resulting in a 70% training set, a 15% validation set and a 15% testing set.

In the data pre-processing stage, we processed the audio and video streams separately. For the video stream, we use the OpenFace toolkit to perform automated face detection, 68 keypoint localisation and face alignment for each video frame, and uniformly scale the aligned face images to a size of 224 × 224 pixels. For the audio stream, we downsampled it from the original 48kHz to 16kHz, and after strict alignment with the video stream, the frames were split using a 25ms Hamming window and a 10ms frame shift. For each audio frame, we extracted features containing the 13-dimensional static MFCC and its first- and second-order differences (39 dimensions in total), along with 64-dimensional log-Meier spectrogram features, and stitched the two together to form a 103-dimensional joint audio feature vector.

In this experiment, weighted accuracy (WA) and weighted F1-score (WF1) were used as the core assessment metrics. WA mitigates the evaluation bias caused by category imbalance by calculating the weighted average of the accuracy in each category (the weight is the true sample size of the category), while F1-score is the reconciled average of precision and recall, and the weighted average version of which comprehensively reflects the overall performance of the model in each category.

In order to fairly verify the advantages of the models in this paper, we choose several representative state-of-the-art models as baselines for comparison:

- Concatenated feature LSTM (CF-LSTM): a simple late fusion baseline. The temporal information of the audiovisual features is extracted using Bi-LSTM respectively, and the final hidden states of the two modalities are spliced at the end of the sequence and fed into the classifier.
- Tensor fusion network (TFN): a classical multimodal fusion model based on tensor outer products that explicitly models inter- and intra-modal interactions.
- Late fusion LSTM (LF-LSTM): audio-visual features are spliced at each time step and fed into the LSTM for time-step-by-time-step fusion.
- Memory fusion network (MFN): a fusion model that utilises multi-view recurrent networks and dynamic memory networks to capture long temporal dependencies across modalities.
- Recurrent attended variation embedding network (RAVEN): a model that employs an attentional mechanism for cross-modal alignment and fusion.

All models are run on the same training, validation, and test sets using the Adam optimiser with an early-stopping strategy to prevent overfitting and ensure fairness in comparisons.

4.2 Results and analysis

We have compared the performance of the proposed model with the above baseline model on the test set and the results are shown in Table 1.

As can be clearly seen from the table, the cross-modal attention fusion-based model proposed in this paper achieves the best performance in both metrics, with a WA of 79.4% and a weighted F1 score of 0.782. Compared with the most powerful baseline model, RAVEN, our model achieves an improvement of 1.4 percentage points in accuracy and 0.013 in F1 score, respectively. This improvement, while not huge, is a significant and practically meaningful advancement on a public dataset that has been extensively studied and is close to saturation in performance, and it validates the effectiveness of our proposed fusion mechanism.

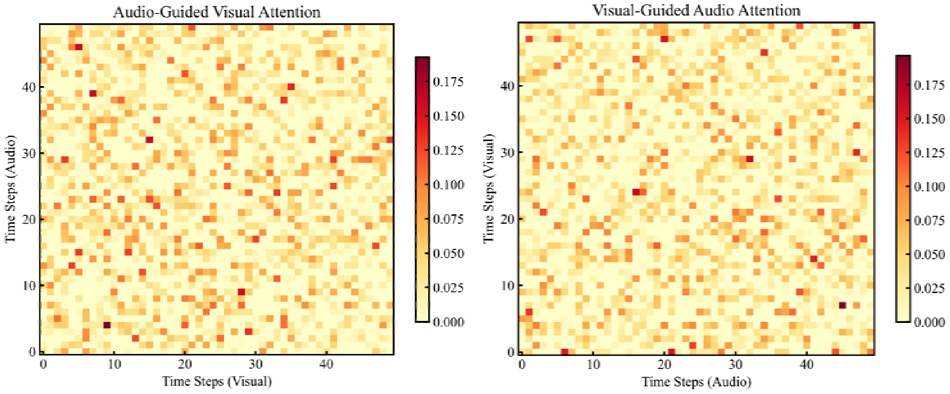
Table 1 Model performance comparison results

Mould	*WA*	*WF1*
CF-LSTM	73.8%	0.728
LF-LSTM	75.1%	0.742
TFN	76.5%	0.758
MFN	77.2%	0.763
RAVEN	78.0%	0.769
Ours	79.4%	0.782

By analysing the baseline model, we can identify some valuable trends: simple splicing fusion has the lowest performance, which illustrates that simple and brute force feature merging is not an optimal solution in multimodal learning. Time-step-by-time-step fusion and tensor fusion with explicitly modelled interactions bring steady performance gains. The models that introduced the memory mechanism and attention mechanism performed even better, demonstrating the importance of dynamic, selective information fusion. It is on this basis that our model is designed with a more refined and bi-directional cross-modal attention-guiding mechanism, which achieves a further breakthrough in performance.

To gain a deeper understanding of the model's decision-making behaviour, we visualise the cross-modal attentional weights of a test sample, as shown in Figure 2. The sample shows a performance of the emotion 'anger'. As can be observed from the heatmap, the model assigns higher weights to the actor's grim facial expression (visual information) at moments when the actor raises the pitch and increases the tone of his voice (audio information is salient), and vice versa for moments when the actor makes a specific angry expression (visual information is salient), the model focuses on acoustic features at that moment accordingly. This synergistic pattern of 'back-and-forth' attention is a vivid illustration of the model's ability to capture fine-grained cross-modal correlations, rather than static averaging.

Figure 2 Heat map of cross-modal attention weights (see online version for colours)



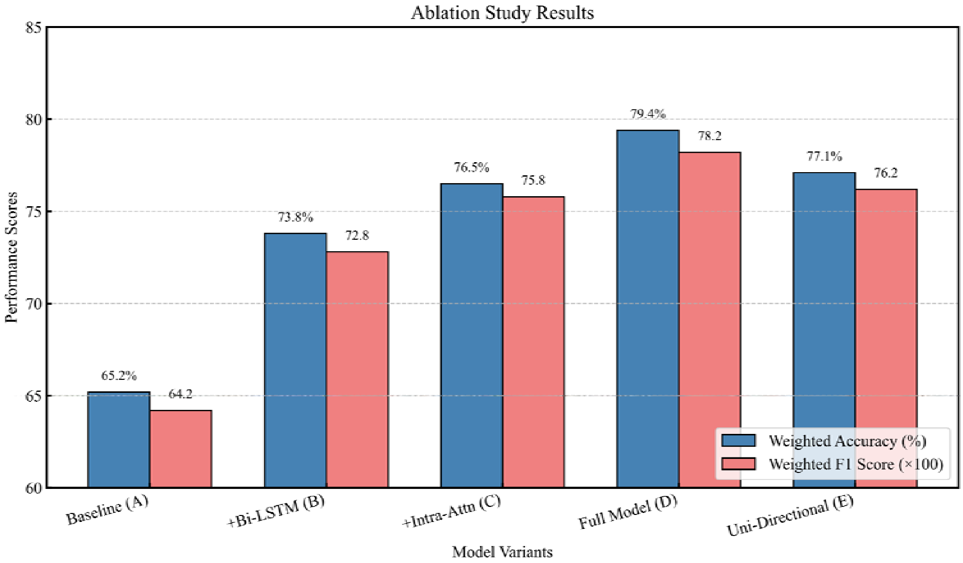
4.3 Ablation experiment

In order to verify the necessity of each key component in the model and its contribution, we designed an exhaustive ablation experiment (Ablation Study). The results are shown in Table 2 and Figure 3.

Table 2 Quantitative comparison of the performance of each model variant of the ablation experiment

Model variant	*WA*	*WF1*
Model A	65.2%	0.642
Model B	73.8%	0.728
Model C	76.5%	0.758
Model D	77.1%	0.762
Model E	79.4%	0.782

Figure 3 Histogram of ablation experiments (see online version for colours)



We first constructed a baseline model A: only frame-level features not processed by Bi-LSTM are used and directly classified after splicing. Subsequently, we add components step by step:

- Model B: on the basis of A, Bi-LSTM timing modelling modules are added for both modalities separately. The results show a substantial improvement in both WA and WF1, which fully demonstrates the extreme importance of modelling temporal context for emotion recognition.
- Model C: based on B, the IMA mechanism is added. The performance is further slightly improved, showing that focusing on the keyframe information inside each modality is effective.

- Model D: i.e., our complete model with the introduction of the IMA fusion module IMA on top of C. The performance of the model D is improved by the addition of the Inter-modal Attention fusion module IMA. The performance reaches the peak among all ablation variants, which strongly demonstrates that the core innovation of our model – the use of inter-modal guidance for information complementarity – is the key to improving performance.

In addition, we tested model E: based on B, using only unidirectional cross-modal attention (audio-only guided vision, or vision-only guided audio). Its performance was significantly lower than that of the full bidirectional model D, which illustrates the reciprocal nature of intermodal influences in emotional expression, where bidirectional guidance is necessary and justified by design.

The results of these ablation experiments are presented in the form of bar charts, which clearly demonstrate the incremental performance improvement brought about by the inclusion of each technique from the baseline to the full model, thus systematically validating the rationality of the model design and the effectiveness of each module.

4.4 Model decision interpretability analysis

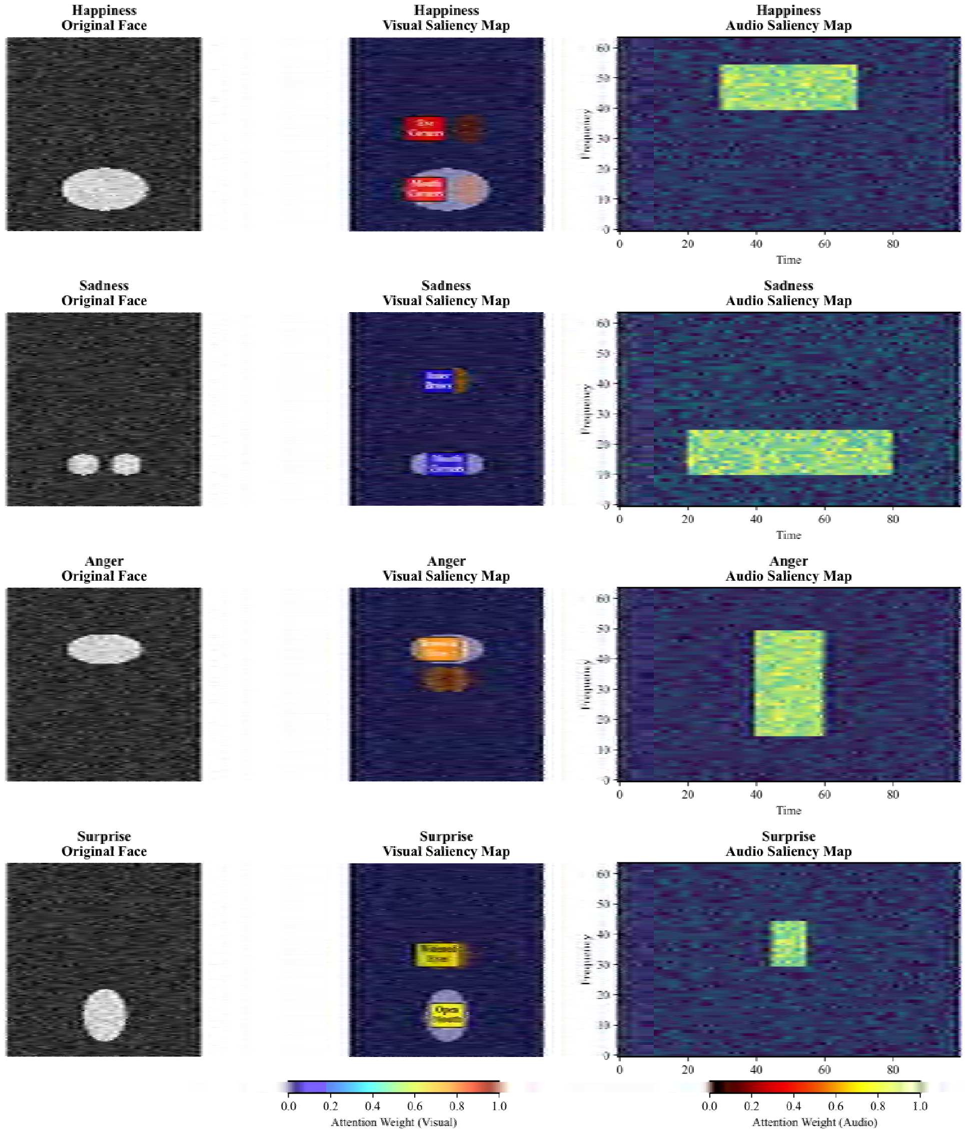
Although deep learning models have achieved excellent performance in emotion recognition tasks, their decision-making process is usually regarded as a ‘black box’, which limits their application in scenarios requiring high confidence, such as educational assessment and clinical assistance. In order to reveal the internal working mechanism of the model and verify whether its decision-making is consistent with human cognition, this experiment introduces the gradient-weighted class activation mapping (Grad-CAM) technique to visualise and analyse the decision-making basis of the model.

We used the Grad-CAM technique to generate saliency maps for audiovisual modalities. For visual modalities, we compute the gradient of the target emotion category with respect to the last convolutional layer feature map and generate a heat map by weighted combination, which highlights the image regions that contribute most to the model decision. For audio modalities, we treat their time-frequency spectrograms as two-dimensional images and use a similar approach to generate saliency maps that identify the temporal and frequency components that are most critical for sentiment classification.

Multiple samples of four typical emotions (happiness, sadness, anger, and surprise) were randomly selected from the test set, and their audiovisual saliency maps were generated and displayed overlaid with the original data for qualitative analysis.

The visualisation results are shown in Figure 4, which provides us with a window into the ‘thinking’ of the model. On the visual side, the saliency map clearly shows that the model’s attention is highly focused on the facial organs that are most relevant to the expression of emotions when judging emotions. For example, in the happy emotion sample, the model focused significantly on the corners of the mouth and eyes, which is highly consistent with the way humans judge happy emotions through smiles and crow’s feet. In the sample of angry emotions, the model’s attention, on the other hand, was focused on the eyebrow, eye, and nose regions, which are precisely the key areas of the human face that exhibit angry features such as frowning and glaring. For the sadness emotion, the model focused on the medial uplift of the eyebrows and the downward pull of the corners of the mouth. And in the emotion of surprise, wide-open eyes and open mouth became the main basis for the model's decision. This highly coincident pattern of attention is strong evidence that our model is not making judgments by memorising irrelevant features of the dataset, but has actually learned to extract biologically meaningful emotional features that are consistent with human cognition.

Figure 4 Model decision significant plot visualisation (see online version for colours)



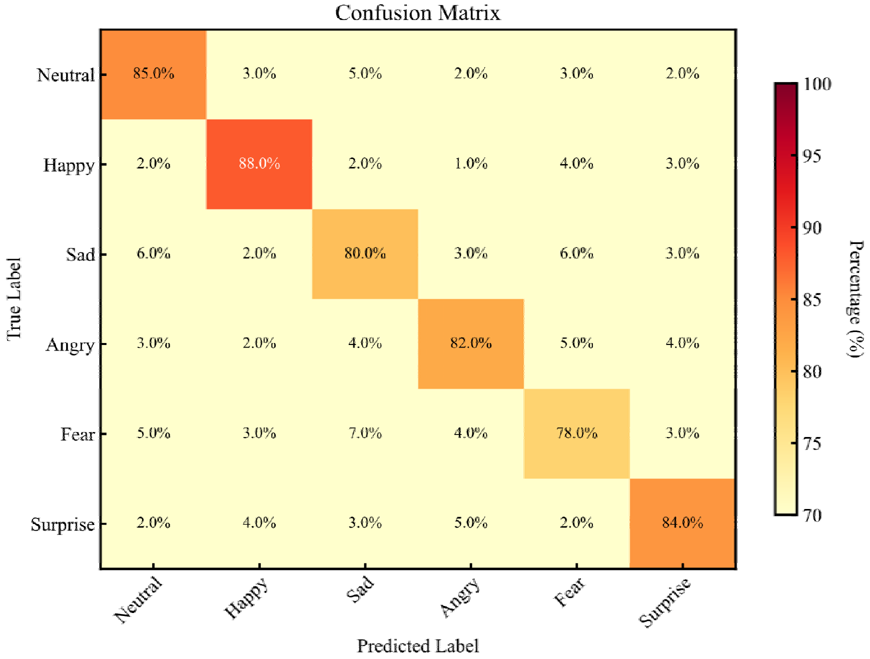
On the audio side, saliency analysis of temporal spectrograms shows that the model is able to acutely capture key acoustic events. For example, in the emotions of anger and surprise, the model assigns high weights to sudden high-energy transients (e.g., bursts of sound, high pitch), while in the emotion of sadness, the model pays more attention to sustained resonance in low-frequency regions and slow intonation changes. This suggests that the audio branch of the model is effective in recognising acoustic attributes associated with emotions.

This interpretability analysis experiment provides an intrinsic, cognitive science-compliant explanation for the model's superior quantitative performance through a qualitative visualisation approach. The results show that the multimodal fusion model proposed in this paper is not only an efficient prediction tool, but also an intelligent system capable of understanding emotional information in a manner consistent with human intuition. This interpretability greatly enhances the credibility of the model and lays a solid foundation for its safe and reliable application in real-world sensitive scenarios.

4.5 Discussion and error analysis

Despite the superior performance achieved by the model in this paper, we have analysed its error cases in depth. Figure 5 shows the confusion matrix of the model on the six categories of emotions. It can be found that the most significant confusions of the model occur between 'fear' and 'sadness', 'anger' and 'surprise', and 'anger' and 'surprise'. between 'fear' and 'sadness', 'anger' and 'surprise'. This is consistent with the common sense of human perception: both 'fear' and 'sadness' may be characterised by a furrowed brow in facial expression and a trembling tone in acoustic features; while strong 'anger' and 'surprise' may be characterised by a frown in facial expression. Both 'anger' and 'surprise' may be characterised by widening of the eyes and raising of the volume of the voice. These cases of confusion illustrate the challenging nature of emotion recognition in children's drama.

Figure 5 Confusion matrix (see online version for colours)



True \ Predicted	Neutral	Happy	Sad	Angry	Fear	Surprise
Neutral	85.0%	3.0%	5.0%	2.0%	3.0%	2.0%
Happy	2.0%	88.0%	2.0%	1.0%	4.0%	3.0%
Sad	6.0%	2.0%	80.0%	3.0%	6.0%	3.0%
Angry	3.0%	2.0%	4.0%	82.0%	5.0%	4.0%
Fear	5.0%	3.0%	7.0%	4.0%	78.0%	3.0%
Surprise	2.0%	4.0%	3.0%	5.0%	2.0%	84.0%

The limitations of the model are mainly in two aspects: first, its dependence on pre-trained face detection and feature extraction models, and the reliability of visual features decreases under extreme lighting, large occlusions, or non-frontal postures. Second, current models mainly focus on transient correlations between tones and facial expressions, and their ability to comprehend rhythms rhythms and narrative contexts on longer temporal sequences remains to be explored.

In summary, this chapter fully demonstrates the effectiveness and sophistication of the cross-modal attention fusion model proposed in this paper through rigorous experimental comparisons, detailed ablation studies, and in-depth error analysis. The experimental results not only show its improvement in quantitative indexes, but also reveal the rationality of its inner working mechanism through visualisation means, which provides a useful reference and a solid foundation for subsequent research.

5 Conclusions

In this paper, a deep learning model based on cross-modal attention fusion is proposed for the emotion recognition task in children's drama performance scenes. The core of this study is to make full use of the complementary information between audio and visual modalities in dramatic performances, and to cope with the challenges posed by children's exaggerated and variable emotional expressions through a refined fusion mechanism.

First, a dual-stream temporal feature extraction network is designed in this paper. On the visual side, a pre-trained deep convolutional network is used to extract frame-level facial features, and a bi-directional long and short-term memory network (Bi-LSTM) is utilised to capture the dynamic evolution of expressions. On the audio side, MFCC and log-Mel spectral features are comprehensively extracted and their long time-series dependencies are modelled by Bi-LSTM as well. This move lays the foundation of high-quality features for subsequent fusion.

Second, the core contribution of this paper is to propose a hierarchical cross-modal attention fusion module. This module not only reinforces the key frame information within each modality through an IMA mechanism, but more importantly introduces a bidirectional cross-modal attention mechanism. The mechanism is guided by the global contextual information of one modality and dynamically and selectively focuses on the most relevant temporal fragments in the other modality, thus realising the adaptive fine-grained inter-modal fusion and generating a unified representation rich in complementary information.

Experimental results show that on the dramatised speech subset of the publicly available dataset RAVDESS, this paper's model achieves a WA of 79.4% with an F1 score of 0.782, which outperforms a variety of state-of-the-art baseline models and validates the effectiveness of the proposed fusion strategy. An exhaustive ablation study further confirms that each component of the model, Bi-LSTM, within-modality attention, and cross-modality attention, are all key to performance improvement. Visual analysis of the confusion matrix and attention weights shows that the model is able to capture cross-modal interaction patterns that are consistent with human cognition, and also reveals that confusing emotion pairs such as 'fear-sadness' and 'anger-surprise' are the key challenges to be solved in the future. The model proposed in this paper has achieved good results.

Although the model proposed in this paper has achieved good results, there are still some limitations that need to be further studied in future work. First, the current model mainly relies on facial expressions and acoustic features, and in the future, multimodal information such as body movements and script context can be incorporated to build a more comprehensive emotion understanding framework. Second, the performance of the model relies to some extent on the accuracy of the face detection and pre-training models, and its robustness in dealing with complex scenes such as extreme poses, occlusion, or low-lighting needs to be strengthened. Finally, future work can explore more explanatory fusion mechanisms and try to apply the model to real educational or medical intervention scenarios for clinical validation.

In summary, this paper provides an effective solution for children's dramatic emotion recognition by innovatively constructing a cross-modal attention fusion model, which provides a useful reference for the application and development of multimodal affective computing in this field.

Acknowledgements

This work is supported by the 2024 Hunan Provincial Social Science Fund Youth Project 'Research on the Innovative Path of Children's Drama in Hunan' (No. 24YBQ131).

Declarations

All authors declare that they have no conflicts of interest.

References

Brauwers, G. and Frasincar, F. (2021) 'A general survey on attention mechanisms in deep learning', IEEE Transactions on Knowledge and Data Engineering, Vol. 35, No. 4, pp. 3279-3298.

Castro, W., Oblitas, J., Santa-Cruz, R. and Avila-George, H. (2017) 'Multilayer perceptron architecture optimization using parallel computing techniques', *PloS One*, Vol. 12, No. 12, p.e0189369.

Chen, Q., Huang, G. and Wang, Y. (2022) 'The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, No. 1, pp.2689–2695.

Choi, H., Cho, K. and Bengio, Y. (2018) 'Fine-grained attention mechanism for neural machine translation', *Neurocomputing*, Vol. 284, No. 1, pp.171–176.

DeRose, J.F., Wang, J. and Berger, M. (2020) 'Attention flows: analyzing and comparing attention mechanisms in language models', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 27, No. 2, pp.1160–1170.

Dong, Y., Jing, C., Mahmud, M., Ng, M.K-P. and Wang, S. (2024) 'Enhancing cross-subject emotion recognition precision through unimodal EEG: a novel emotion preceptor model', *Brain Informatics*, Vol. 11, No. 1, p.31.

Ghaffarian, S., Valente, J., Van Der Voort, M. and Tekinerdogan, B. (2021) 'Effect of attention mechanism in deep learning-based remote sensing image processing: a systematic literature review', *Remote Sensing*, Vol. 13, No. 15, p.2965.

Graves, A. and Schmidhuber, J. (2005) 'Framewise phoneme classification with bidirectional LSTM and other neural network architectures', *Neural networks*, Vol. 18, Nos. 5–6, pp.602–610.

Guo, M-H., Xu, T-X., Liu, J-J., Liu, Z-N., Jiang, P-T., Mu, T-J., Zhang, S-H., Martin, R.R., Cheng, M-M. and Hu, S-M. (2022) 'Attention mechanisms in computer vision: a survey', *Computational Visual Media*, Vol. 8, No. 3, pp.331–368.

Hernández, A. and Amigó, J.M. (2021) 'Attention mechanisms and their applications to complex systems', *Entropy*, Vol. 23, No. 3, p.283.

Jia, N., Zheng, C. and Sun, W. (2022) 'A multimodal emotion recognition model integrating speech, video and MoCAP', *Multimedia Tools and Applications*, Vol. 81, No. 22, pp.32265–32286.

Kheradpisheh, S.R., Ganjtabesh, M., Thorpe, S.J. and Masquelier, T. (2018) 'STDP-based spiking deep convolutional neural networks for object recognition', *Neural Networks*, Vol. 99, No. 1, pp.56–67.

Ku, H. and Dong, W. (2020) 'Face recognition based on MTCNN and convolutional neural network', *Frontiers in Signal Processing*, Vol. 4, No. 1, pp.37–42.

Lange, M. and Scheve, C.V. (2021) 'Valuation on financial markets: calculations of emotions and emotional calculations', *Current Sociology*, Vol. 69, No. 5, pp.761–780.

Li, D., Liu, J., Yang, Z., Sun, L. and Wang, Z. (2021) 'Speech emotion recognition using recurrent neural networks with directional self-attention', *Expert Systems with Applications*, Vol. 173, No. 1, p.114683.

Li, J., Jin, K., Zhou, D., Kubota, N. and Ju, Z. (2020) 'Attention mechanism-based CNN for facial expression recognition', *Neurocomputing*, Vol. 411, No. 1, pp. 340–350.

Li, S., Zhang, T. and Chen, C.P. (2024) 'Sia-net: Sparse interactive attention network for multimodal emotion recognition', *IEEE Transactions on Computational Social Systems*, Vol. 11, No. 5, pp.6782–6794.

Li, W., Liu, K., Zhang, L. and Cheng, F. (2020) 'Object detection based on an adaptive attention mechanism', *Scientific Reports*, Vol. 10, No. 1, p.11307.

Li, X., Li, M., Yan, P., Li, G., Jiang, Y., Luo, H. and Yin, S. (2023) 'Deep learning attention mechanism in medical image analysis: basics and beyonds', *International Journal of Network Dynamics and Intelligence*, Vol. 10, No. 1, pp.93–116.

Liu, J-W., Liu, J-W. and Luo, X-L. (2021) 'Research progress in attention mechanism in deep learning', *Chinese Journal of Engineering*, Vol. 43, No. 11, pp.1499–1511.

Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X. and Zheng, W. (2023) 'The multi-modal fusion in visual question answering: a review of attention mechanisms', *PeerJ Computer Science*, Vol. 9, No. 1, p.e1400.

Lv, H., Chen, J., Pan, T., Zhang, T., Feng, Y. and Liu, S. (2022) 'Attention mechanism in intelligent fault diagnosis of machinery: a review of technique and application', *Measurement*, Vol. 199, No. 1, p.111594.

Nguyen, M.T., Lin, W.W. and Huang, J.H. (2023) 'Heart sound classification using deep learning techniques based on log-Mel spectrogram', *Circuits, Systems, and Signal Processing*, Vol. 42, No. 1, pp.344–360.

Niu, Z., Zhong, G. and Yu, H. (2021) 'A review on the attention mechanism of deep learning', *Neurocomputing*, Vol. 452, No. 1, pp.48–62.

Pan, J., Fang, W., Zhang, Z., Chen, B., Zhang, Z. and Wang, S. (2023) 'Multimodal emotion recognition based on facial expressions, speech, and EEG', *IEEE Open Journal of Engineering in Medicine and Biology*, Vol. 5, No. 1, pp.396–403.

Rabinovich, M., Tristan, I. and Varona, P. (2013) 'Neural dynamics of attentional cross-modality control', *PloS One*, Vol. 8, No. 5, p.e64406.

Rodriguez, P., Velazquez, D., Cucurull, G., Gonfaus, J.M., Roca, F.X. and Gonzalez, J. (2019) 'Pay attention to the activations: A modular attention mechanism for fine-grained image recognition', *IEEE Transactions on Multimedia*, Vol. 22, No. 2, pp.502–514.

Soydaner, D. (2022) ‘Attention mechanism in neural networks: where it comes and where it goes’, *Neural Computing and Applications*, Vol. 34, No. 16, pp.13371–13385.

Suebsombut, P., Sekhari, A., Sureephong, P., Belhi, A. and Bouras, A. (2021) ‘Field data forecasting using LSTM and Bi-LSTM approaches’, *Applied Sciences*, Vol. 11, No. 24, p.11820.

Wei, Z., Cui, Z. and Zeng, J. (2012) ‘Social emotional optimisation algorithm with emotional model’, *International Journal of Computational Science and Engineering*, Vol. 7, No. 2, pp.125–132.

Zhang, S., Yang, Y., Chen, C., Liu, R., Tao, X., Guo, W., Xu, Y. and Zhao, X. (2023) ‘Multimodal emotion recognition based on audio and text by using hybrid attention networks’, *Biomedical Signal Processing and Control*, Vol. 85, No. 1, p.105052.