SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 989.
https://doi.org/10.62617/mcb989

Article

Emotion detection in artistic creation: A multi-sensor fusion approach leveraging biomechanical cues and enhanced CNN models

Peng Du

School of Art Education, Hubei Institute of Fine Arts, Wuhan 430205, China; 18611411580@163.com

CITATION

Du P. Emotion detection in artistic creation: A multi-sensor fusion approach leveraging biomechanical cues and enhanced CNN models. Molecular & Cellular Biomechanics. 2025; 22(4): 989. https://doi.org/10.62617/mcb989

ARTICLE INFO

Received: 4 December 2024
Accepted: 26 February 2025
Available online: 13 March 2025

Abstract: Artistic creation is a means of expressing human emotions. To intuitively capture the emotions conveyed by the artist in their works, we propose an improved CNN-based emotion detection method that incorporates biomechanical elements. Recognizing that emotions are accompanied by physiological and biomechanical responses such as heart rate variations, facial muscle activity, and speech tone fluctuations, we collect and integrate multi-sensor data, including heart rate, facial expression, and verbal expression. This information is processed through a multi-sensor signals fusion method based on an enhanced Convolutional Neural Networks (CNN), which allows for the extraction of rich and accurate emotional feature representations from the creator's biomechanical signals. In particular, the facial muscle movements and subtle variations in speech tone, which are integral to understanding emotional states, are effectively captured and analyzed. Furthermore, we introduce a Conditioning Diffusion Model for Emotion Prediction, where emotional features, informed by biomechanical responses, serve as semantic conditions to boost the accuracy of emotion detection. This approach enables precise identification of the artist's emotions by considering the intricate interplay of physiological and biomechanical signals. Experimental results demonstrate that our proposed method achieves an mAP score of 85.36%, an MSE score of 0.73%, and a runtime of 87 milliseconds, providing technical support for predicting the emotions of creators based on their biomechanical responses.

Keywords: artistic creation; biofeedback analysis; emotion detection; CNN

1. Introduction

Since ancient times, artistic creation has been a key way to express human emotions. With the changes of The Times and the progress of society, the expression of emotions and ideas in artistic creation has attracted more and more attention from the public and has become more and more integrated with aesthetic emotions. The emotional state of artists in creation is full of appeal, which reflects their aesthetic perspective on the creation object. This kind of creative emotion, which is different from everyday emotion, is rooted in life experience and presents typical characteristics after refining. With the help of modern technologies such as artificial intelligence and big data, real-time monitoring and analysis of artists' emotions [1,2] can provide precise guidance for creation. At the same time, in-depth exploration of emotions in art creation [3] can help artists integrate emotions more skillfully into their works, thereby improving the artistic charm of their works.

At present, the research of emotion detection faces multiple technical challenges. The first problem is the subjectivity of emotion, which leads to different people's emotional judgments of the same text, speech or image, and then leads to inconsistent labels and semantic ambiguity [4,5]. Secondly, the complexity and diversity of emotions increase the difficulty of detection because the boundaries between different

emotions are fuzzy, such as pleasure and happiness, and there are also differences in the degree within the same emotion type, such as the subtle difference between happiness and excitement [6]. Moreover, single-modal data cannot fully capture emotional information, and it is necessary to integrate multi-modal features such as vision, speech, and text [7]. However, there are semantic differences between different modalities, which increases the difficulty of integration [8,9]. Finally, the emotion prediction results need to be interpretable to enhance users' understanding of the reliability and confidence of the model. However, current machine-learning techniques have limitations in explaining the decision-making process of emotion detection [10].

To solve the above problems, many emotion detection models have been proposed.

Tzirakis et al. [11] utilized ResNet to individually achieve audio and visual features. Then, the features were connected for fusion and input into a two-layer LSTM to predict valence and arousal. In addition, attentions are applied to emotion recognition (ER) based on visual and auditory fusion. Zhao et al. [12] proposed an attention to achieve emotional regions from visual and auditory modalities, respectively. Hao et al. [13] suggested an integrated vision-audio ER framework, which uses an SVM and the convolutional neural network (CNN) with deep-learned visual and audio features to generate four sub-models and then fuses the results of these sub-models to predict emotions, achieving decision-level fusion. Li et al. [14] suggested the parallel structure for the speech ER network, which augments the global acoustic features with local spectral details of the whole speech to achieve accurate detection and recognition of human voice emotion. Li et al. [15] suggested the multi-modal method based on the graph and attention to achieve the interplay of information across different modalities for emotion detection in conversation. Wei et al. [16] constructed a dataset to analyze the emotion using EEG and physiological signals. They monitored the EEG, surrounding physiological data, and positive facial videos of 32 subjects while watching 40 one-minute music video samples and used the subjects' evaluations after watching the videos as labels. The researchers used the bionic grey wolf algorithm to identify the emotional valence and arousal. To study the emotional features expressed by the joint visual channel of the face and body, Wei et al. [17] established a bimodal database containing facial and body gestures, designed an algorithm and a bimodal classifier to simultaneously analyze the kinematic features of both, and realized the accurate detection and recognition of six types of emotions. Han et al. [18] proposed a fuzzy logic-based multimodal fusion network that operates in a multi-feature space and is specifically designed for emotion recognition in bimodal tasks involving facial expressions and orchestra conducting gestures.

Although the current models for emotion detection in artistic creation have achieved remarkable results, they still face many challenges, such as the difficulty of representing emotional features in artistic works, and the effective fusion and reasoning of multimodal features. Given the limitations of traditional emotion recognition methods, particularly those that solely rely on facial expression or speech analysis, which are susceptible to interference from subjects' subjective factors, leading to inaccurate and unreliable emotion assessment results, we have decided to adopt a comprehensive approach that integrates biomechanical analysis [19,20] for

assessing individuals' emotional states. This integrated emotion detection technique, which leverages biomechanical analysis, focuses on precisely determining emotional states by analyzing physiological signals (such as heart rate variability) generated during emotional changes. These physiological signal changes are spontaneously regulated by the autonomic nervous system, free from the influence of individuals' subjective consciousness, thereby providing a more objective and authentic basis for emotion assessment [21]. Furthermore, this method boasts non-invasiveness and high accuracy. Therefore, we propose an art-creation emotion detection method based on improved CNN. This method combines the advanced technology of deep learning with the in-depth understanding of art theory, aiming to capture and analyze the complex emotions in art creation more accurately. Our model first utilizes the powerful feature extraction ability of CNN to extract low-level feature representations from different modal information of artworks (such as images, text, audio, etc.). These features not only include basic visual elements, text words or audio waveforms, but also cover deeper semantic and emotional cues. Subsequently, we design a multi-modal fusion module, which adopts the visual diffusion mechanism to dynamically adjust the importance of different modal features to achieve effective information integration. Through this mechanism, the model can focus on the modal information that is most critical for emotion expression, while suppressing the interference of noise and secondary information, thereby improving the accuracy and robustness of emotion detection.

2. Related works

In art creation, many achievements of emotion detection and recognition based on machine learning have been born.

In single-modal emotion analysis, Jelodar et al. [22] utilized the LSTM method to classify the emotions of COVID-19 texts on social media, which is significant to understand the attitude and research of social issues. Zhang et al. [23] leveraged part-of-speech rules to discern various product attributes, grounded in fine-grained sentiment analysis and the Kano model, which avoids omissions caused by multi-word over-segmentation and the gap caused by sentiment analysis [24] and need identification. Lou et al. [25] retrieve sentiment information and syntactic information of sentences from external commonsense knowledge, construct a sentiment graph and dependency graph for each sentence, and then propose a sentiment dependency graph convolutional network framework, which has excellent results in the sarcasm detection. Hassan et al. [26] trained a deep visual sentiment analysis model to analyze opinions on natural disasters and images on social platforms. Yadav et al. [27] suggested a novel network with multi-scale features for sentiment classification, which combines different levels of deep representation to classify the sentiment of images.

For multimodal emotion analysis, Alfreihat et al. [28] proposed a text-image neighborhood binary classifier based on text features. Experiments show that it has better performance than using only text for sentiment classification. Khan et al. [29] suggested a two-stream framework, which applied the Transformer to translate images, then utilized a non-autoregressive text generation approach with the single channel to extract text features, and then used the translated form to construct auxiliary sentences

to provide multimodal information for the language model. This model combines visual information and linguistic information to improve the accuracy. Zhu et al. [30] suggested a new network to relate the emotional image regions to texts [31] and achieved excellent results. Liang et al. [32] determined the interactions between image and text by detecting regions in images and designed a cross-modal GCN to understand the coordination relationship between modalities in detection.

3. Artistic emotion detection method based on improved CNN

Aiming at the complexity of emotion feature representation and the challenge of effective fusion of multimodal features, we design an artistic creation emotion detection scheme based on improved CNN. This scheme optimizes the fusion process of multi-sensor signal features through the CNN framework and constructs a correlation model between multi-modal emotional factors, which includes the following two parts: First, the improved CNN is used to realize the efficient fusion of multi-sensor signal features; secondly, the diffusion model is used to construct the emotional factor model to enhance the ability to capture and analyze emotional features.

Firstly, we design a multi-sensor signal fusion method based on improved CNN, which can efficiently fuse the signal features from different sensors. By simulating the propagation process of information in the sensor network, the model can capture the potential connections and mutual influence between signals to generate more rich and accurate emotional feature representations. This fusion strategy not only improves the dimension and depth of emotional features but can also enhance the model to understand complex emotions.

Then, to establish the relationship between multimodal emotional factors, we introduce the diffusion model. This model can automatically adjust its structure and parameters according to the dynamic changes of emotional features to accurately capture the complex associations between emotional factors. By constructing the graph structure between emotional factors, we obtain the comprehensive analysis of emotional states. This adaptivity enables our model to better adapt to the diversity and uncertainty of emotional states in the process of artistic creation.

3.1. Multi-sensor signals fusion method based on improved CNN

In the process of artistic creation, the author's emotion is the key point that affects his or her results. We use a variety of sensors to collect features that can reflect the creator's emotions, such as heart rate (ECG), electromyography (EMG), facial expression (FE), kinematic data (KD), language expression (LE), and pressure data (PD). At the same time, we propose a multi-sensor signal feature fusion method based on improved CNN (MSF).

For the extraction of cardiac rhythm signals, we used the way of ECG acquisition. The ECG waveform is composed of key components such as the *P* wave, QRS complex, and *T* wave, which collectively record the electrical activity of the heart over a while. A typical periodic ECG signal waveform is shown in **Figure 1**. To accurately identify the key location of the ECG signal and determine the duration of the *P* wave, QRS complex, and *T* wave, we used the modified wavelet transform method. The core

of this method lies in the design of a special wavelet filter. The coefficients of the high-pass filter are set as (3/4, 1/2, 1/2, 3/4), while the coefficients of the low-pass filter are set as (−1/4, 3/4, 3/4, −1/4). Among them, the continuous wavelet transform is the core of the filter, and its formula is

$$\chi(u,v) = \int x(t)\psi(t-vu)dt \tag{1}$$

where χ denotes the wavelet coefficient and x is the original signal, namely ECG; ψ refers to the wavelet basis function, we use the Haar function; u is the scale parameter and v is the translation parameter. Then, we look for the extreme points by checking the detail coefficients at each scale and comparing them to 0. These extremal points provide us with clues to the location of key ECG features. Based on these extreme points, we set appropriate thresholds to accurately identify and locate *P* waves, QRS complexes, and *T* waves. After successfully locating these waveforms, we further extracted five key feature vectors. These feature vectors include PR interval, ST interval, *P* wave, *T* wave and *R* wave.

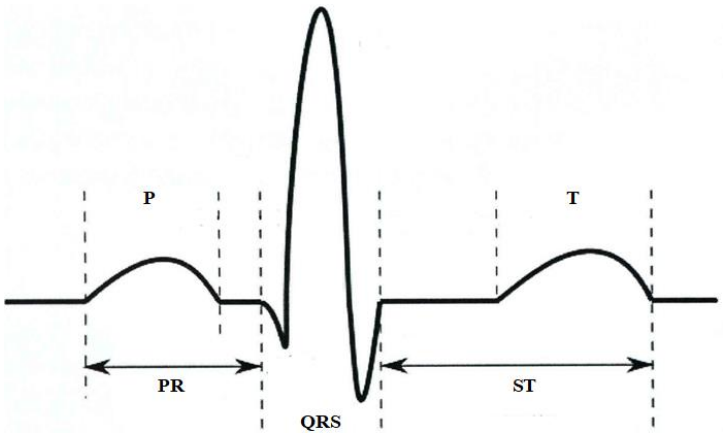

Figure 1. ECG waveform diagram.

To capture facial emotional expressions, we use conventional camera devices to directly capture image information. Based on this image data, we develop an optimized multi-scale convolutional network architecture for processing facial images, whose specific structure is shown in **Figure 2**. The processing flow first involves designing a diverse grid partition to decompose the image into grid features containing information at different scales. Then, these grid features are integrated and deeply fused by applying 7 × 7, 3 × 3, and 1 × 1 kernels to achieve the emotional features from the creator's face. The convolution is calculated by Equation (2):

$$O(x,y) = \sum_{m=0}^{i-1}\sum_{n=0}^{j-1} I\left(x+m-\frac{i-1}{2}, y+n-\frac{j-1}{2}\right)\cdot K(m,n) \tag{2}$$

where $O(x, y)$ denotes the corresponding output pixel, $I(,)$ denotes the pixel value at the corresponding position of the input image, $K(m, n)$ refers to the value of the kernel K, i and j denote the size of the convolution kernel. We can not only capture detailed information through the multi-scale convolutional network architecture, which is suitable for extracting small-scale or fine-grained emotional features, but also high-level features that are rich in semantic information, aiding in the recognition of large-

scale or holistic features. Furthermore, this architecture can extract contextual information surrounding the features, and the incorporation of contextual information provides clues for feature recognition in complex emotional scenarios.

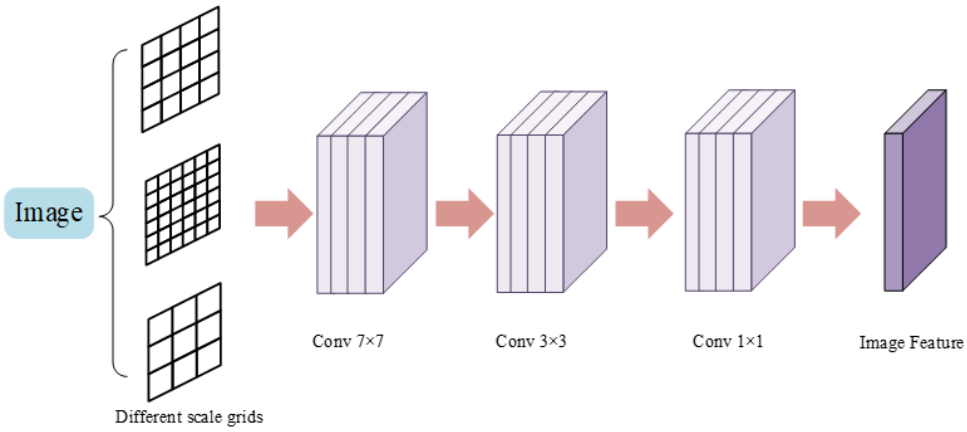

Figure 2. Improved multi-scale convolutional network.

Considering the temporal nature of EMG signals, linguistic content, kinematic data, and pressure data, we integrate CNN with Transformers, as illustrated in **Figure 3**. To extract multi-level semantic information from these data, we employ convolutional layers with kernel sizes of 3, 5, and 7 to separately capture the features of the four types of data. These extracted features are then concatenated along the feature channel dimension. To fully comprehend the semantic relationships within the context, we further utilize an LSTM network to extract semantic features, yielding an initial representation that encapsulates the entire linguistic expression. To enable the model to focus on critical regions of the speech features, we apply an attention mechanism to further process the features, guiding the model to attend to important semantic content. This allows the model to capture muscle contraction levels in the arms and fingers, track movement patterns of wrist and finger joints, and measure contact pressure between creative tools and media.

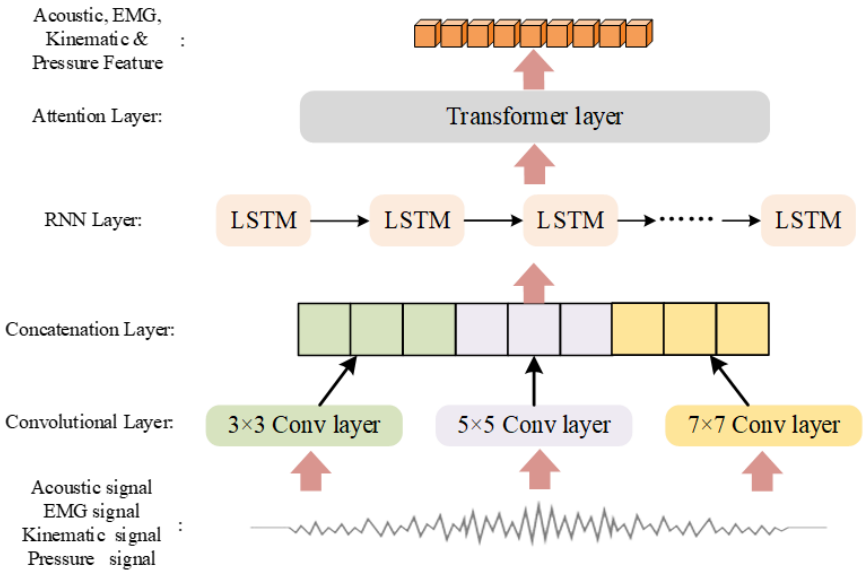

Figure 3. Feature extraction network for language expression content.

3.2. Conditioning diffusion model for emotion prediction

To fully leverage the benefits of ECG, facial expression, and language expression features in the emotion prediction task of creators, a conditioning diffusion model for emotion prediction (CDM) is proposed, and the framework is in **Figure 4**. This method effectively integrates the information of these three modalities through an innovative fusion strategy, with the goal of enhancing the precision and dependability of emotion prediction.

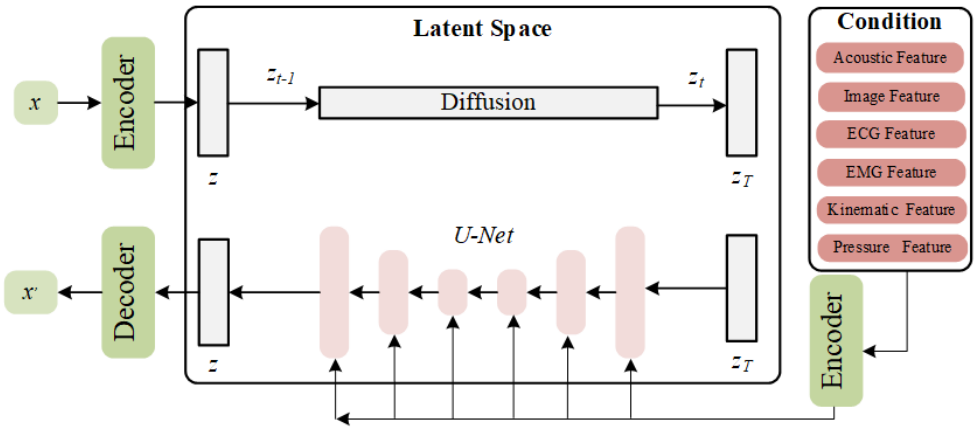

Figure 4. Emotion prediction method based on a diffusion model.

According to ECG, facial expression and language expression, we regard them as three independent additional semantic conditions, and build an emotion prediction model based on them. This model can effectively predict the emotional state of the creator by relying on the learned emotional feature patterns. As for the diffusion model, its core formula elaborates how to generate new data samples by gradually adding noise during the forward diffusion process. The formula is as follows:

$$x_t = \sqrt{a_t} \cdot x_{t-1} + \sqrt{1 - a_t} \cdot \epsilon_t \tag{3}$$

where the new data sample generated at step t of x_t, at refers to a gradually decreasing parameter in the range [0, 1], controlling the degree of noise added. x_{t-1} denotes the data sample at step $t - 1$, which is the data sample from the previous step. ϵ_t denotes standard Gaussian noise, which follows the standard normal distribution.

Our emotion prediction model learns the complex relationship between the given three features and the emotional state of the creator by analyzing them. These features may include text content, speech features, social media behaviors, etc., which each carry different semantic information and together constitute a comprehensive description of the creator's emotion. Through the training process, the model learns to extract key emotional feature patterns from the input features, which can reflect the emotional state of the creator in different contexts. Once the model is trained, it can receive new feature inputs and predict the current or future emotional state of the creator. This predictive ability is of great significance for understanding the psychological state of creators, optimizing user experience, and providing personalized services.

4. Experiment and analysis

4.1. Dataset and implementation settings

We used a multimodal emotion dataset (https://zenodo.org/records/13717256, doi: 10.5281/zenodo.13717256 and https://zenodo.org/records/10584026, doi: 10.5281/zenodo.10584026) of the art emotion prediction model for testing. The dataset comprises paired painting-audio sets categorized according to five distinct emotions: anger, sadness, neutrality, fun, and happiness. Additionally, it includes over 23,500 sentence utterance videos sourced from more than 1000 YouTube speakers. These sentences are randomly selected from topics and monologue videos.

During the training phase, we utilized a Ryzen 9700x processor coupled with four Nvidia RTX 4090 GPUs to boost computational efficiency. To speed up the process, we chose Pytorch as our framework and carefully adjusted its configurations to align precisely with the parameters detailed in **Table 1**. We undertake the following preprocessing procedures for multimodal data: We tackle missing data by employing techniques like imputation, deletion of affected records, or prediction of missing values. To balance the sample sizes across different categories in the dataset, we employ resampling methods. Furthermore, we augment data diversity and generalization capacity by creating visual data samples via techniques such as rotation, flipping, scaling, and cropping. Additionally, we apply filters and other methods to purify the data, removing noise from both language and ECG signals.

Table 1. Detail settings.

Parameters	**value**
Learning rate	2×10^{-4}
Epoch	40
Dropout	0.75
Layer number	12

To thoroughly assess the predictive performance, we have selected the mean square error (MSE), mean average precision (mAP), precision (*P*), recall (*R*), and *F*-measure as evaluation metrics for each model. The corresponding formulas are outlined below:

$$P = \frac{Num(gt \cap pr)}{Num(pr)} \tag{4}$$

$$R = \frac{Num(gt \cap pr)}{Num(gt)} \tag{5}$$

$$F = \frac{2 \times (P \times R)}{P + R} \tag{6}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{gt} - \hat{y}_{pr})^2 \tag{7}$$

$$mAP = \frac{1}{N} \times \sum P_n \times R_n \tag{8}$$

where Num(.) is used for the quantity, *gt* refers to the ground truth, pr denotes the prediction results, and *N* is the total number of samples. Besides, we also apply costing time and visualization to evaluate the method while comparing it with others.

4.2. Ablation experiments

Firstly, we delve into the impact of MSF (and CDM) on sentiment prediction model performance. By executing meticulously planned ablation experiments, we thoroughly assess the performance of each module, utilizing the ResNet101 architecture as our benchmark model. The experimental results are clearly presented in **Table 2**. The data reveals that the MSF method alone can elevate the *F*-score by 2.41%, while the CDM method independently achieves a more substantial *F*-score increase of 4.37% compared to the baseline. Notably, when MSF and CDM are combined, the emotion prediction model undergoes further enhancement, with the mAP reaching 85.36% and the MSE dropping to 0.73%, fully showcasing the robust potential of their collaborative efforts. The MSF method significantly bolsters the model's capacity to capture intricate emotional information by effectively amalgamating signal features from diverse sensors. This approach leverages the complementary strengths of different sensors in capturing emotion-related signals, thereby mitigating the information bias or noise potentially introduced by a single sensor. By integrating MSF, the model gains a more comprehensive understanding of the intrinsic attributes of emotional data. On the other hand, the CDM method further optimizes the model for sentiment prediction tasks by incorporating the conditional diffusion mechanism. This method taps into the latent distribution characteristics of sentiment data and captures nuanced changes in sentiment by mimicking the diffusion process. This mechanism allows for precise delineation of boundaries between different emotional states, demonstrating heightened sensitivity and accuracy in predictions. The complementarity between MSF and CDM is fully harnessed. MSF furnishes the model with comprehensive and abundant emotional information, while CDM further refines the classification and recognition of emotions on this foundation. This synergy not only elevates the model's overall performance but also renders it more adaptable and precise in navigating complex emotional scenarios.

Table 2. Ablation experiments for FEF and ALF.

MSF	CDM	*P* (%)	*R* (%)	*F* (%)	mAP (%)	MSE (%)
		81.32	75.68	78.24	80.94	2.64
√		82.65	78.48	80.65	83.87	1.41
	√	83.26	79.12	82.61	84.59	1.22
√	√	85.53	81.02	83.24	85.36	0.73

Then, based on fixed CDM, we further explored the influence of the features of ECG, FE and LE (including electromyography, kinematic data, language expression and pressure data during entire experiments) on the performance of the model. Through comparative analysis, as shown in **Figure 5** and **Table 2**, it can be observed

that either feature alone can enhance the mode. This finding indicates that each feature contains unique information that positively contributes to model performance. At the same time, when we try to mix the three features of ECG, FE and LE, the model is further significantly improved. This improvement is not only reflected in the improvement of accuracy, but also in the recognition and processing ability of the model for complex situations. This result strongly proves the performance of feature fusion and the significant synergistic effect of ECG, FE and LE features in improving the performance of the model. We can conclude that rationally using and fusing multiple features such as ECG, FE and LE is an effective solution to improve the model.

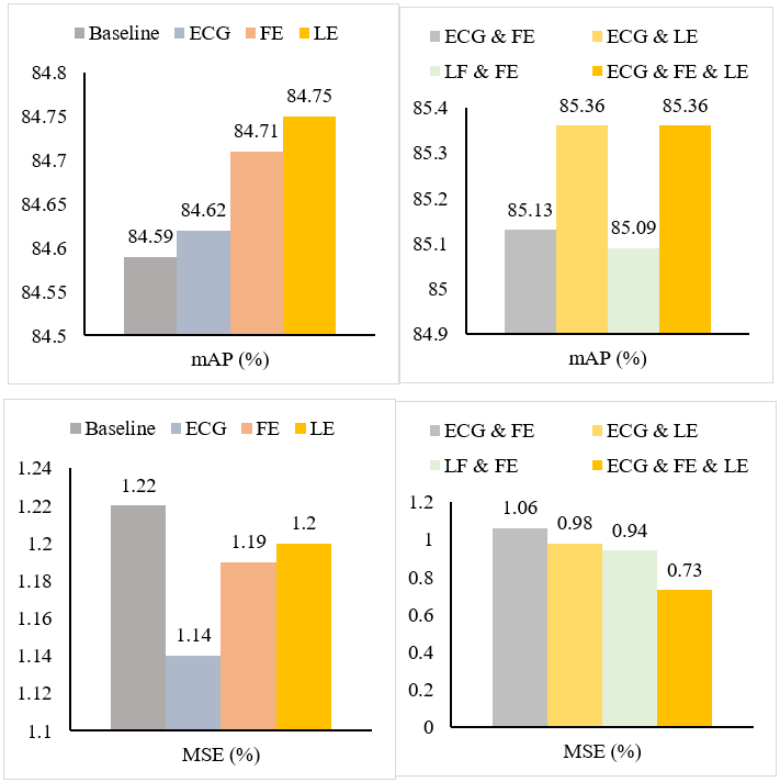

Figure 5. The ablation of ECG, FE and LE while applying CDM.

When exploring the relationship between creators' emotional states and their biomechanical features, the two distinct mood states of positive and negative emotions have become the focus of our research. Through detailed analysis, we found that ECG, EMG, pressure data, and motion data exhibit different features under these two emotional states. When creators are in a positive emotional state, their biomechanical features demonstrate remarkable stability and coordination. Firstly, in terms of heart rate variability, a positive emotion results in a stable and regular heart rate, indicating that the creator's heart function remains stable under pleasant or exciting emotions, providing adequate blood supply and oxygen to the body and ensuring the physiological foundation during the creative process. Simultaneously, positive emotions lead to smoother and more rhythmic muscle contractions. Further analysis of EMG signals clearly shows that muscle activation patterns are more orderly under positive emotions. Additionally, the analysis of kinematic and pressure features also reveals the positive impact of positive emotions on the stability of creators' movements.

In a positive emotional state, parameters such as hand movement speed, acceleration, and tremor frequency exhibit stability and regularity, and a stable level of pressure control is maintained. In contrast to positive emotions, negative emotions have a significant negative impact on creators' biomechanical features. In a negative emotional state, creators often experience increased heart rate variability; muscle tension increases, and contraction patterns become more disrupted. The analysis of kinematic and pressure features similarly reveals the destructive effect of negative emotions on the stability of creators' movements.

4.3. Compare other methods

We perform an in-depth and comprehensive performance evaluation of the newly proposed emotion prediction model, aiming to confirm its effectiveness and superior performance in real-world application scenarios. To this end, we carefully selected the methods form Zhang et al., [33], Nie et al., [34], Sharma et al., [35], Gupta [36], Kumar et al., [37], and Joseph et al., [38] as the comparison benchmarks, which are extremely representative and cutting-edge research results in this field. **Table 3** compares experiments for our method.

Table 3. Ablation experiments for ECG, FE and LE.

ECG	FE	LE	CDM	*P* (%)	*R* (%)	*F* (%)
			√	83.26	79.12	82.61
√			√	84.36	80.31	82.88
	√		√	83.26	80.53	82.69
		√	√	83.11	79.54	81.75
√	√		√	84.65	80.19	82.64
√		√	√	84.79	80.03	82.97
	√	√	√	84.23	81.02	82.19
√	√	√	√	85.53	81.02	83.24

Based on the results presented in **Table 4** and **Figure 6**, our sentiment prediction model excels across all evaluation metrics. Specifically, the model achieves an MSE as low as 0.73%, an mAP as high as 85.36%, a *P*-score of 85.53%, an *R*-score of 81.02%, and an F1-score of 83.24%. Compared to the selected baseline models, our method demonstrates significant superiority in performance. Even when compared to other competitive methods in the field, our method maintains a lead in accuracy and recall rates by at least 1.41% and 0.57%, respectively. Our method boasts an MSE that is 0.16% lower than that of Paper [38], an mAP that is 2.71% higher than that of Paper [37], and an mAP that is 1.47% higher than that of Paper [36]. These data fully demonstrate the high efficiency and advanced nature of our sentiment prediction model in practical applications.

Table 4. Comparison with others in terms of mAP and MSE for ECG, FE and LE.

Methods	mAP (%)	MSE (%)
Baseline	80.94	2.64
Paper [33]	83.21	1.26
Paper [34]	84.10	1.54
Paper [35]	82.35	0.88
Paper [36]	83.89	0.97
Paper [37]	82.65	1.21
Paper [38]	84.68	0.89
Ours	85.36	0.73

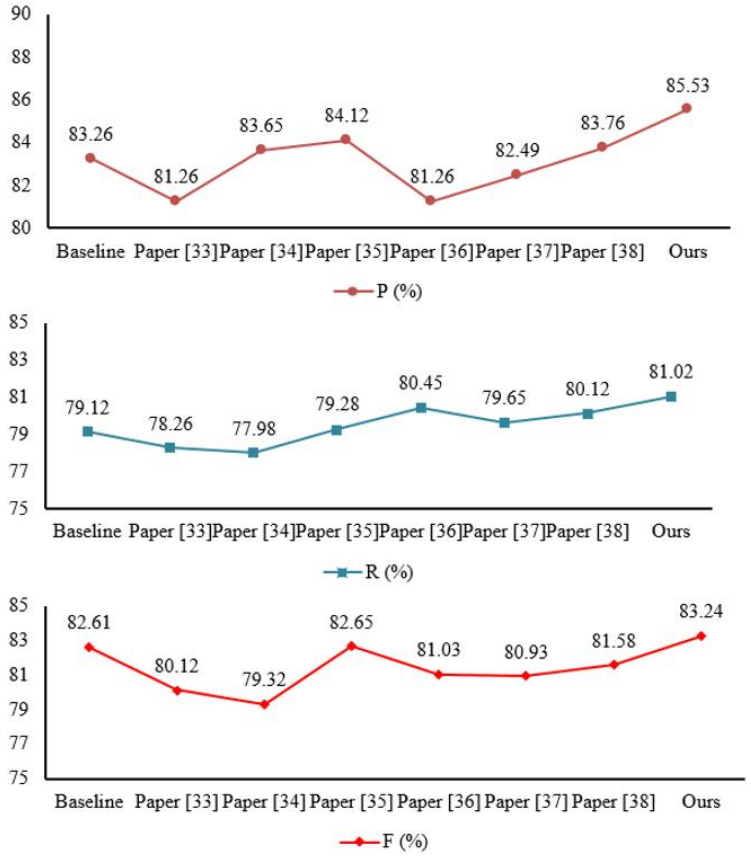

Figure 6. Comparison results in terms of *P*, *R* and *F*.

ECG, FE and LE features each play an indispensable role in explaining the excellent performance of our emotion prediction model. ECG features reflect the physiological response of emotions by capturing subtle changes in cardiac activity. These features can not only help the model understand the physiological stress response in emotional states but also provide immediate feedback on emotional changes. In our model, ECG features play a key role in improving the low value of the MSE metric, which means that the model has higher accuracy in predicting emotions. FE features provide rich emotional cues to the model by capturing and analyzing the movement changes of facial muscles. These features not only cover basic emotional expressions (e.g., happiness, sadness, anger, etc.) but also capture more subtle emotional changes (e.g., surprise, contempt, etc.). In our model, FE features are

beneficial to improving mAP and *P* scores. LE features reveal the linguistic features of emotional expressions by analyzing lexical, grammatical, and semantic information in texts. These features can not only help the model understand the emotional tendency in the text but also capture the emotional color and context information behind the text. Through deep analysis of text data, the model can accurately identify and understand the emotional information in language expressions, thereby improving the coverage and comprehensive performance of emotion prediction.

4.4. Application testing

Firstly, to comprehensively measure the efficiency and practicality of our method in the emotion prediction task, we use a confusion matrix, a visual tool, to show the specific results of each emotion detection. The confusion matrix is the visualization tool in the performance evaluation of supervised learning algorithms. As shown in **Figure 7**, through the confusion matrix, we can conclude that the prediction accuracy of our method is over 82% on each emotion class.

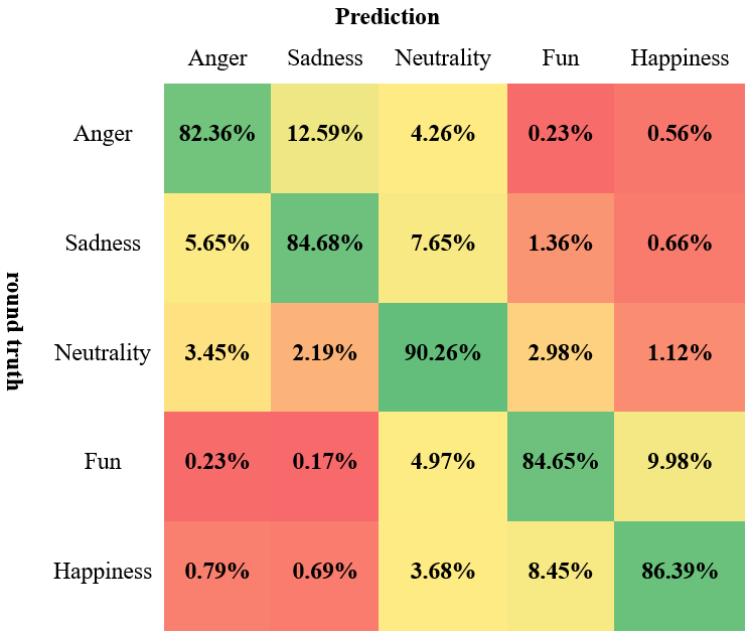

Figure 7. Our method's emotion prediction confusion matrix.

Further analyzing the details of the confusion matrix, we find that for "neutral" emotion, our method shows extremely high recognition ability, with an accuracy of more than 90%. In the prediction of the "sadness" emotion, although the accuracy is slightly lower than that of the "happy" emotion, it still maintains at 84.68%, which reflects the strong recognition sensitivity of the model to negative emotion. For the more intense emotions like anger, the prediction accuracy of the model is 82.36%, indicating that although the model faces certain challenges in distinguishing these emotions, it can still maintain high accuracy. We also observe some specific error patterns in the confusion matrix, such as misclassifying "funny" as "happy" or "sad" as "neutral." These errors suggest that future model optimization can further enhance the ability to learn and distinguish these confusing emotional features. Emotion is essentially a continuous variation of features, and its complexity and diversity are

difficult to fully capture through simple discretization. When using convolutional neural networks (CNNs) for emotion recognition, the tendency of CNNs to discretize continuous data may fail to adequately reflect the subtle changes and diversity of emotions, thereby increasing the error rate in emotion recognition. Additionally, in multi-sensor emotion recognition tasks, the amount of information provided by different modalities is often uneven, and the noise levels within each modality also vary. This imbalance in information and the presence of noise challenge the fusion of multi-sensor signals. During the fusion process, the risk of information loss or misdirection increases, further affecting the accuracy of emotion recognition. More importantly, when CNNs handle multi-sensor signal fusion, their inherent feature extraction capabilities may not be sufficient to account for multi-dimensional feature information. The expression of emotions often involves multiple aspects of features, such as changes in facial expressions, tone of voice, speaking speed, and physiological signal variations. If CNNs cannot effectively capture and integrate this multidimensional feature information, the accuracy of emotion recognition will be compromised, potentially leading to incorrect recognition results. Therefore, to improve the accuracy of emotion recognition, we need to explore more advanced algorithms and technologies to better handle continuous emotion features while effectively fusing multi-modal information and considering multi-dimensional feature extraction. By doing so, we can more accurately understand and recognize emotions, providing stronger support for the application of artificial intelligence in the field of affective computing.

Then, we test the emotional changes of an artist during a complete creative process and highlight the proposed method through a sequential comparison of different methods. The sample is a pessimistic artwork, which takes a total of two hours to create. We recorded the mood every 10 min and used our method to predict the mood. The results are shown in **Figure 8**, where we used 1 to 5 to represent five different emotions: anger, sadness, neutrality, fun, and happiness. We can conclude that our method is the closest to the true value of the sample.

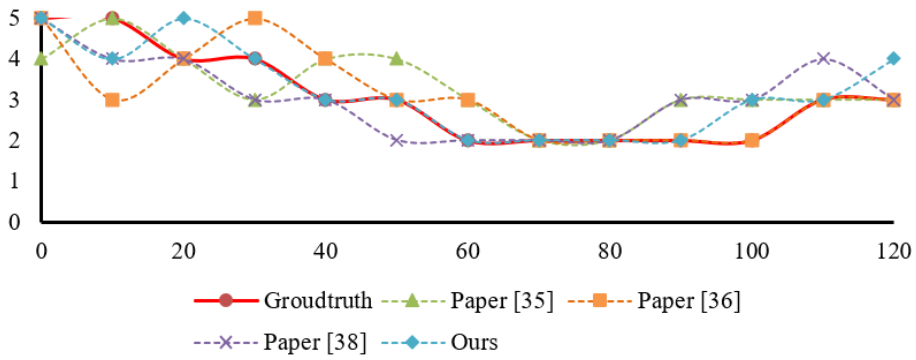

Figure 8. Comparison of emotions from a pessimistic art creator.

Finally, we evaluated the execution time of our approach, the results of which are shown in **Figure 9**. Through observation, it can be found that our method is in the middle level of all tested methods in terms of time consumption. However, given its actual performance in prediction performance, such time consumption is completely acceptable.

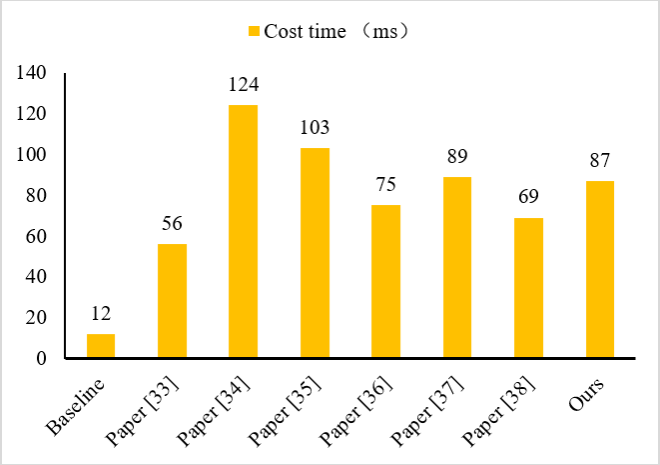

Figure 9. Comparison of cost time of our method and others.

4.5. Discuss

After experimental verification, we tested the multi-sensor signal fusion method based on improved CNN and combined it with the conditioning diffusion model for emotion prediction. The experimental data fully show that our proposed model has shown high effectiveness and wide practicability in the creators' emotion prediction task.

By fusing data from different sensors, such as ECG, facial expressions and language expressions, the model can comprehensively and deeply capture the emotional changes of creators. The improved CNN structure enables the model to efficiently extract the key features in this complex data, and the conditional diffusion model further enhances the ability of the model to predict the dynamic changes of emotions.

In the experiments, we use a diverse dataset, including creator samples under different emotional states, to ensure the generalization ability of the model. Through comparative experiments, we find that compared with traditional emotion prediction methods, our model shows significant advantages in prediction accuracy, robustness and real-time performance. In particular, when dealing with complex emotional states, such as mixed emotions or subtle emotional changes, our model shows higher sensitivity and accuracy. In addition, we also evaluate the computational efficiency and resource consumption. The results demonstrate that though the model has achieved significant performance improvement, its computational complexity and resource requirements do not increase significantly, which is significant for the practical applications.

Our proposed multi-sensor signal fusion method based on improved CNN combined with the emotion prediction model of the conditional diffusion model not only shows high efficiency and practicality in the author's emotion prediction task but also provides strong technical support for the further development of the emotional intelligence field.

5. Conclusion

To directly experience the author's emotion in the process of artistic creation, we propose an artistic creation emotion detection method for improved CNN. By extracting the information of heart rhythm, facial expression and language expression, a multi-sensor signals fusion method based on improved CNN is constructed to fuse and obtain the multi-modal emotional feature representation. The multimodal features are used as semantic conditions to construct a conditioning diffusion model, which enhances emotion detection and realizes the accurate recognition of artistic creation. Experiments show that the *P*-value of 85.53%, the *R*-value of 81.02 and the *F*-value of 83.24 are obtained by our proposed method, which can realize the emotion recognition and analysis of artistic creation. By integrating biomechanical analysis methods, this study establishes a mapping relationship between emotional states and biomechanical features during the artistic creation process, providing new insights for emotion detection research. Future research will explore the biomechanical feature differences across various art forms and their influence on emotional expression.

Acknowledgments: The author would like to thank the anonymous reviewers for their valuable comments on this paper.

Availability of data and materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Liu Z, Zhang T, Yang K, et al. Emotion detection for misinformation: A review. Information Fusion. 2024; 107: 102300. doi: 10.1016/j.inffus.2024.102300
2. Nie W, Bao Y, Zhao Y, et al. Long Dialogue Emotion Detection Based on Commonsense Knowledge Graph Guidance. IEEE Transactions on Multimedia. 2024; 26: 514-528. doi: 10.1109/tmm.2023.3267295
3. Mamieva D, Abdusalomov AB, Kutlimuratov A, et al. Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. Sensors. 2023; 23(12): 5475. doi: 10.3390/s23125475
4. Shin H, Lee B, Ku B, et al. Noisy label facial expression recognition via face-specific label distribution learning. Image and Vision Computing. 2024; 143: 104901. doi: 10.1016/j.imavis.2024.104901
5. Hung LP, Alias S. Beyond Sentiment Analysis: A Review of Recent Trends in Text Based Sentiment Analysis and Emotion Detection. Journal of Advanced Computational Intelligence and Intelligent Informatics. 2023; 27(1): 84-95. doi: 10.20965/jaciii.2023.p0084
6. Khan M, El Saddik A, Alotaibi FS, et al. AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network. Knowledge-Based Systems. 2023; 270: 110525. doi: 10.1016/j.knosys.2023.110525
7. Min C, Lin H, Li X, et al. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. Information Fusion. 2023; 96: 214-223. doi: 10.1016/j.inffus.2023.03.015
8. Nandini D, Yadav J, Rani A, et al. Design of subject independent 3D VAD emotion detection system using EEG signals and machine learning algorithms. Biomedical Signal Processing and Control. 2023; 85: 104894. doi: 10.1016/j.bspc.2023.104894
9. Krishnamoorthy P, Sathiyanarayanan M, Proença HP. A novel and secured email classification and emotion detection using hybrid deep neural network. International Journal of Cognitive Computing in Engineering. 2024; 5: 44-57. doi: 10.1016/j.ijcce.2024.01.002

10. Oğuz FE, Alkan A, Schöler T. Emotion detection from ECG signals with different learning algorithms and automated feature engineering. Signal, Image and Video Processing. 2023; 17(7): 3783-3791. doi: 10.1007/s11760-023-02606-y
11. Tzirakis P, Zhang J, Schuller BW. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018.
12. Zhao S, Ma Y, Gu Y, et al. An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos. Proceedings of the AAAI Conference on Artificial Intelligence. 2020; 34(01): 303-311. doi: 10.1609/aaai.v34i01.5364
13. Hao M, Cao WH, Liu ZT, et al. Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. Neurocomputing. 2020; 391: 42-51. doi: 10.1016/j.neucom.2020.01.048
14. Li W, Xue J, Tan R, et al. Global-Local-Feature-Fused Driver Speech Emotion Detection for Intelligent Cockpit in Automated Driving. IEEE Transactions on Intelligent Vehicles. 2023; 8(4): 2684-2697. doi: 10.1109/tiv.2023.3259988
15. Li J, Wang X, Lv G, et al. GA2MIF: Graph and Attention Based Two-Stage Multi-Source Information Fusion for Conversational Emotion Detection. IEEE Transactions on Affective Computing. 2024; 15(1): 130-143. doi: 10.1109/taffc.2023.3261279
16. Wei D, Chen D, Huang Z, et al. An improved chaotic GWO-LGBM hybrid algorithm for emotion recognition. Biomedical Signal Processing and Control. 2024; 98: 106768. doi: 10.1016/j.bspc.2024.106768
17. Wei J, Hu G, Yang X, et al. Learning facial expression and body gesture visual information for video emotion recognition. Expert Systems with Applications. 2024; 237: 121419. doi: 10.1016/j.eswa.2023.121419
18. Han X, Chen F, Ban J. FMFN: A Fuzzy Multimodal Fusion Network for Emotion Recognition in Ensemble Conducting. IEEE Transactions on Fuzzy Systems. 2025; 33(1): 168-179. doi: 10.1109/tfuzz.2024.3373125
19. Mahfoudi MA, Meyer A, Gaudin T, et al. Emotion Expression in Human Body Posture and Movement: A Survey on Intelligible Motion Factors, Quantification and Validation. IEEE Transactions on Affective Computing. 2023; 14(4): 2697-2721. doi: 10.1109/taffc.2022.3226252
20. Straker R, Exell TA, Farana R, et al. Biomechanical responses to landing strategies of female artistic gymnasts. European Journal of Sport Science. 2021; 22(11): 1678-1685. doi: 10.1080/17461391.2021.1976842
21. Coombes SA, Higgins T, Gamble KM, et al. Attentional control theory: Anxiety, emotion, and motor planning. Journal of Anxiety Disorders. 2009; 23(8): 1072-1079. doi: 10.1016/j.janxdis.2009.07.009
22. Jelodar H, Wang Y, Orji R, et al. Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. IEEE Journal of Biomedical and Health Informatics. 2020; 24(10): 2733-2742. doi: 10.1109/jbhi.2020.3001216
23. Zhang J, Zhang A, Liu D, et al. Customer preferences extraction for air purifiers based on fine-grained sentiment analysis of online reviews. Knowledge-Based Systems. 2021; 228: 107259. doi: 10.1016/j.knosys.2021.107259
24. Meng J, Dong Y, Long Y, et al. An attention network based on feature sequences for cross-domain sentiment classification. Intelligent Data Analysis. 2021; 25(3): 627-640. doi: 10.3233/ida-205130
25. Lou C, Liang B, Gui L, et al. Affective Dependency Graph for Sarcasm Detection. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021.
26. Hassan SZ, Ahmad K, Hicks S, et al. Visual Sentiment Analysis from Disaster Images in Social Media. Sensors. 2022; 22(10): 3628. doi: 10.3390/s22103628
27. Yadav A, Vishwakarma DK. A Deep Multi-level Attentive Network for Multimodal Sentiment Analysis. ACM Transactions on Multimedia Computing, Communications, and Applications. 2023; 19(1): 1-19. doi: 10.1145/3517139
28. Alfreihat M, Almousa OS, Tashtoush Y, et al. Emo-SL Framework: Emoji Sentiment Lexicon Using Text-Based Features and Machine Learning for Sentiment Analysis. IEEE Access. 2024; 12: 81793-81812. doi: 10.1109/access.2024.3382836
29. Khan Z, Fu Y. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021.
30. Zhu T, Li L, Yang J, et al. Multimodal Sentiment Analysis With Image-Text Interaction Network. IEEE Transactions on Multimedia. 2023; 25: 3375-3385. doi: 10.1109/tmm.2022.3160060
31. Das R, Singh TD. Image–Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. ACM Transactions on Asian and Low-Resource Language Information Processing. 2023; 22(6): 1-30. doi: 10.1145/3584861
32. Liang B, Lou C, Li X, et al. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022.

33. Zhang Z, Wang L, Yang J. Weakly Supervised Video Emotion Detection and Prediction via Cross-Modal Temporal Erasing Network. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023.
34. Nie L, Li B, Du Y, et al. Deep learning strategies with CReToNeXt-YOLOv5 for advanced pig face emotion detection. Scientific Reports. 2024; 14(1). doi: 10.1038/s41598-024-51755-8
35. Sharma S, S R, Akhtar MdS, et al. Emotion-Aware Multimodal Fusion for Meme Emotion Detection. IEEE Transactions on Affective Computing. 2024; 15(3): 1800-1811. doi: 10.1109/taffc.2024.3378698
36. Gupta BB, Gaurav A, Chui KT, et al. Deep Learning-Based Facial Emotion Detection in the Metaverse. In: Proceedings of the 2024 IEEE International Conference on Consumer Electronics (ICCE); 2024.
37. Kumar A. A systematic analysis of machine learning algorithms for human emotion detection using facial expression. International conference on signal processing & communication engineering systems: SPACES-2021. 2024; 2512: 020021. doi: 10.1063/5.0112473
38. Joseph A, Carvalho S, Saldanha N, et al. Emotion Detection Based on Text and Emojis. In: Proceedings of the 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS); 2024.